Advances in Cognitive Diagnosis Modeling

by

Miguel A. Sorrel

A doctoral dissertation submitted in partial fulfillment for the degree of Doctor of Psychology

in the

Faculty of Psychology
Universidad Autónoma de Madrid

Dissertation Directors: Prof. Francisco J. Abad and Julio Olea

April 2018

February 19, 2018

This thesis is dedicated to the memory of my grandfather Gervasio Luján,
an extraordinary man whom I still miss every day.

Agradecimientos/Acknowledgements

El pequeño o gran aporte que suponga esta tesis hubiese sido imposible sin la participación de ciertas personas e instituciones que me han acompañado durante mi etapa doctoral.

Gracias principalmente a mis directores, los Dres. Francisco José Abad y Julio Olea, que han sido un ejemplo a seguir tanto profesional como personalmente. Guardo muchos momentos con cariño. Su apoyo y confianza en mi trabajo han sido fundamentales. Intentaré mantener la constancia y la humildad a lo largo de mi carrera. Si consigo parecerme a vosotros ya habré ganado mucho.

A very special gratitude goes out to Dr. Jimmy de la Torre for all his guidance, patience, and academically and emotionally support. You challenged me for the better, and for that I will always be thankful.

Gracias al Dr. Vicente Ponsoda que junto con Jimmy me recibió cuando aún era un estudiante de grado y me permitieron pensar a lo grande. A la Dra. Carmen García por su trato siempre agradable.

A los coautores de los trabajos incluidos en la tesis. Especialmente al Dr. Juan Ramón Barrada al que tuve la ocasión de conocer mejor durante el viaje a Chicago. Many thanks, too, to Dr. Filip Lievens, I really enjoyed the time we spent working together on the revision of the ORM paper.

Gracias a mis compañeros en el pasillo y en el laboratorio 17. Ha sido increíble compartir todos estos momentos juntos. No sólo me llevo compañeros, si no también grandes amigos: David Santos, David Moreno, Blanca Requero y Loli,

Gracias a mis padres, Andrés y María, que son lo mejor de mi vida. A mi hermana Laura que cuida de mí. A mi tía Núria que me hizo creer en grandes expectativas. A mi familia.

Finalmente, gracias a tí, Rocío. Eres mi compañera y la que mejor entiende como me he sentido durante estos años, siendo siempre un apoyo incombustible. No sé si hubiese sido posible sin tí, pero sin duda tú haces que merezca la pena.

Abstract

ADVANCES IN COGNITIVE DIAGNOSIS MODELING

by

MIGUEL A. SORREL

Dissertation Directors: Francisco J. Abad and Julio Olea

Cognitive diagnosis models (CDMs) have shown a rapid development over the past decades. This set of restricted latent class models is the basis for a new psychometric framework where the dimensions underlying performance on the tests are assumed to be discrete. Notwithstanding the progress achieved, some aspects have not been fully explored. It is for this reason that this dissertation aims to contribute in three directions. (1) Broadening the area of application of CDMs. Empirical data is used to illustrate how CDMs provide a new approach that not only overcomes the limitations of the conventional methods for assessing the validity and reliability of situational judgment tests (SJTs) scores, but that also allows for a deeper understanding on what SJTs really measure. The data set comes from an application of a SJT that presents situations about student-related issues. (2) Evaluating item-level model fit statistics. Factors such as generating model, test length, sample size, item quality, and correlational structure are considered in two different Monte Carlo studies. The performance of several statistics and different strategies to cope with poor-quality data are discussed. Additionally, the two-step likelihood ratio test is introduced as a new index for item-level model comparison. (3) Introducing model comparison as a way of improving cognitive diagnosis computerized adaptive testing (CD-CAT) applications. Accuracy and item usage of a CD-CAT based on the combination of models selected with the new item-level model comparison statistic are explored under different calibration sample size, Q-matrix complexity, and item bank length conditions using Monte Carlo methods. The advantages of this approach over the application of a single reduced CDM or a general model are discussed. In general, the results of the studies included in this dissertation can be the basis for more reliable assessments and indicate the importance of selecting an appropriate psychometric framework. Item-level model selection emerges as a new and promising strategy to make the best of our data that can be generalized to other psychometric frameworks such as traditional item response theory.

Resumen

AVANCES EN MODELADO DIAGNOSTICO COGNITIVO

por

MIGUEL A. SORREL

Directores de la tesis: Francisco J. Abad and Julio Olea

Los modelos de diagnóstico cognitivo (MDC) han mostrado un rápido desarrollo en las últimas décadas. Este conjunto de modelos de clase latente restringida es la base de un nuevo marco psicométrico donde se asume que las dimensiones subyacentes al rendimiento en los test se asumen a ser discretas. A pesar del progreso logrado, algunos aspectos no han sido completamente explorados. Es por esta razón que esta tesis pretende realizar contribuciones en tres direcciones. (1) Ampliar el área de aplicación de los MDC. Se emplean datos empíricos para ilustrar cómo los MDC proporcionan un nuevo enfoque que no sólo supera las limitaciones de los métodos convencionales para evaluar la validez y fiabilidad de las puntuaciones obtenidas con test de juicio situacional (TJS), sino que también permite una comprensión más profunda acerca de lo que los TJS realmente miden. La base de datos proviene de una aplicación de un TJS que presenta situaciones sobre problemas relacionados con los estudiantes. (2) Evaluación de los estadísticos de ajuste a nivel de ítem. Factores tales como el modelo generador de los datos, la longitud del test, el tamaño muestral, la calidad de los ítems y la estructura de correlaciones se consideran en dos estudios diferentes de simulación Monte Carlo. Se discute el rendimiento de varios estadísticos y diferentes estrategias para lidiar con datos de baja calidad. Además, la prueba de razón de verosimilitud en dos pasos se presenta como un nuevo índice para la comparación de modelos a nivel de ítem. (3) Introducción de la comparación de modelos como una forma de mejorar las aplicaciones de test adaptativos computarizados de diagnóstico cognitivo (TAI-DC). La precisión y el uso de los ítems de un TAI-DC basado en la combinación de modelos seleccionados con el nuevo estadístico para la comparación de modelos a nivel de ítem se exploran bajo diferentes condiciones de tamaño de la muestra de calibración, complejidad de la matriz Q y longitud del banco de ítems utilizando métodos de simulación Monte Carlo. Se discuten las ventajas de este enfoque sobre la aplicación de un único MDC reducido o un modelo general. En general, los resultados de los estudios incluidos en esta tesis pueden ser la base para evaluaciones más precisas e indican la importancia de seleccionar un marco psicométrico apropiado. La selección de modelos a nivel de ítem surge como una estrategia nueva y prometedora para aprovechar al máximo nuestros datos que puede generalizarse a otros marcos psicométricos, como la teoría de respuesta al ítem tradicional.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation of the dissertation

As a result of the evolution of the psychometric theory, the emergence of new models has been made possible the application of item response theory (IRT) with different response formats (e.g., polytomous, continuous, forced-choice), and tests assessing more than one dimension using multidimensional-IRT and bi-factor modeling. What all the above-mentioned developments have in common is that the underlying latent traits are assumed to be continuous. In the last decades, a new psychometric framework has emerged: the cognitive diagnosis modeling (CDM) framework. This new set of models emerged with the purpose of diagnostically classifying the examinees in a predetermined set of discrete latent traits, typically denoted as attributes. Attributes are discrete in nature rather than continuous, with typically only two levels indicating if the examinees mastered or not mastered each specific attribute. Compared to the large amount of research in the traditional IRT context, only a small number of studies have been conducted in the context of CDM.

Current publications in this field focus on the introduction of new models (e.g., de la Torre and Chiu, 2016; Henson et al., 2009; von Davier, 2005), adapting methodologies from the traditional IRT context (e.g., Cheng, 2009; Cui and Li, 2015; Wang et al., 2015), and new methodologies applicable to CDM (e.g., de la Torre and Chiu, 2016; Kaplan et al., 2015; Kuo et al., 2017). Meanwhile, some review articles describing the state-of-the-art in CDM were published in different international journals and book chapters (Akbay and Kaplan, 2017; DiBello et al., 2007; Huebner, 2010; Rupp and Templin, 2008). The year 2010 was an important year because it was published the book "*Diagnostic Measurement: Theory, Methods, and Applications*", by Rupp, Templin, and Henson. To this day, this is probably the reference manual in CDM. In a first stage, authors shared their codes written in different softwares such as **Mplus** (Muthén and Muthén, 2013), Ox (Doornik and Ooms, 2007), and R (Team, 2016). The publication of the **CDM** (Robitzsch et al., 2017) and **GDINA** (Ma and de la Torre, 2017) R packages made these methodologies much more accessible.

CDM is undoubtedly one of the fashion themes in current psychometrics. For example, this can be noted in a large number of papers and workshops during the last conferences of the National Council on Measurement in Education, American Educational Research Association, and the Psychometric Society. Most presentations introduced new methods for model estimation, model fit, and test structure assessment, and compare existing statistics. Clearly, CDM is a relatively new area of research that has been aided by the experiences in the traditional IRT context. Notwithstanding that, everything suggests that in the coming years we should be able to see how some CDM methodologies are exported to the traditional IRT context (e.g., Magis and Barrada, 2017; Sorrel et al., 2018b). In either case, there are still some aspects that should continue to be worked on in the CDM context. This is the context where this dissertation arises, under the supervision of Francisco José Abad and Julio Olea, directors of the Spanish Ministry of Economy and Competitiveness project entitled "*Computerized adaptive testing based on new psychometric models*" (PSI2013-44300-P).

In the light of the above, and in the development period of CDM, this dissertation aims to contribute in three directions:

1. Broadening the area of application of CDMs. CDM emerged in the area of education as a way to identify students' strengths and weaknesses, and develop remedial instructions (e.g., Tatsuoka, 1983; Haertel, 1989). Later, CDMs were applied to measure psychological disorders (de la Torre et al., 2017; Templin and Henson, 2006). This dissertation resumes and expands the pinoneering study by García et al. (2014) introducing CDMs as a new psychometric framework to evaluate situational judgement tests (SJTs). This is in line with current lines of research in SJT according to the recent review by Weekley et al. (2015). The results of the first study of this dissertation, published in *Organizational Research Methods*, has already had continuity in leading scientific journals (Bley, 2017; Chen and Zhou, 2017).

2. Evaluating item-level model fit statistics. Even though some statistics for assessing absolute and relative fit have already been used in CDM (de la Torre and Lee, 2013; Kunina-Habenicht et al., 2012), some of the classical statistics in traditional IRT have not been evaluated. These include the likelihood ratio (LR) test and Lagrange multiplier (LM) test for assessing relative fit, and the gold standard for absolute fit evaluation, namely the $S - X^2$ introduced by Orlando and Thissen (2000, 2003). In addition, some factors such as item quality have been generally overlooked in the previous simulation studies. Different levels were selected based on a literature review of current CDM applications within and without the educational field.

3. Introducing model comparison as a way of improving adaptive testing applications. CDMs have been recently applied as a basis for adapting testing. Most current research is related to item

selection and stopping rules. A prior concern is item bank calibration. Model selection can be problematic in CDM given the wide range of available models (a detailed review is provided in Subsection 1.2.2). Following the results obtained with fixed test versions (Ma et al., 2016; Rojas et al., 2012), item-level model comparison indices will be introduced as a way of improving accuracy and item usage by selecting the most appropriate model for each item.

The rest of the dissertation is organized as follows. The remaining sections of Chapter 1 describe the goals of the dissertation and state of the art in CDM research in the particular areas that are of interest in this dissertation (i.e., current empirical applications, model fit evaluation, computerized adaptive testing). Chapter 2, Chapter 3, Chapter 4, and Chapter 5 are devoted to the four studies included. Specifically, Chapter 2 introduces the application of CDMs to the area of Industrial-Organizational psychology and SJTs data, describing the sequential steps in the application of CDMs. Chapter 3 and Chapter 4 focus on item-level model fit evaluation. Chapter 3 systematically compares the performance of different inferential item-fit indices, introducing the LR test and the LM test in the context of CDM. Chapter 4 proposes an approximation to the LR test, the 2LR test, that uses a two-step estimation approach to speed up the computation of the LR test. Chapter 5 explores the application of item-level model comparison indices as a previous step in CD-CAT that will serve to improve accuracy and test security. Chapter 6 provides a general discussion of the results obtained in this dissertation, and details the limitations and future research lines. Finally, the list of publications derived from this dissertation, and the Spanish version of the General Discussion can be found in Appendix A and Appendix B.

1.2 An Introduction to cognitive diagnosis models

1.2.1 Cognitive diagnosis modeling

Over the last years, there has been an increase of interest in a group of psychometric models known as CDMs. This new area of research has come to be called *cognitive diagnosis modeling*. Based on the review of existing labels for these models that have been used in the literature (e.g., cognitively diagnostic models, Henson and Douglas 2005; cognitive psychometric models, Rupp and Mislevy 2007; multiple classification models, Haertel 1989; structured IRT models, Rupp and Mislevy 2007), Rupp and Templin (2008) offered the following definition:

> Diagnostic classification models (DCM) are probabilistic, confirmatory multidimensional latent-variable models with a simple or complex loading structure. They are suitable for modelling observable categorical response variables and contain unobservable (i.e., latent) categorical predictor variables. The predictor variables are combined in compensatory and

> non-compensatory ways to generate latent classes. DCM enable multiple criterion-referenced interpretations and associated feedback for diagnostic purposes, which is typically provided at a relatively fine-grain size. This feedback can be, but does not have to be, based on a theory of response processing grounded in applied cognitive psychology. Some DCM are further able to handle complex sampling designs for items and respondents, as well as heterogeneity due to strategy use. (p.226).

According to this definition, cognitive diagnosis modeling is an interdisciplinary approach to diagnostic assessment. The theory on a particular domain is used to model the psychological processes underlying performance on the test items. In this sense, cognitive diagnosis modeling establishes a link between cognitive psychology and statistical modeling.

Early applications of cognitive diagnosis modeling took place in the area of educational measurement (Tatsuoka, 1983; Haertel, 1989). Traditionally, educational assessments have been used to provide a single score that identifies the student location along a single proficiency continuum. The most sophisticated statistical tools used to address this goal were often rooted in traditional IRT, an ideal alternative to overcome some of the limitations of the classical test theory (CTT). One of the major contributions of IRT is the extension of the concept of reliability. Within the IRT framework, precision is not uniform across the entire range of test scores. Both CTT and IRT frameworks were the basis for summative assessments. Summative assessments are given periodically to determine rank-order comparisons among students, or against certain standards. Examples of summative assessments include end of semester ratings, state exams (e.g., Graduate Record Examinations; GRE), and international assessments like The Programme for the International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS). These assessments serve many different purposes, including identifying the level of proficiency and differentiating passing from non-passing students (de la Torre and Minchen, 2014). As such, they fulfill an important function in education.

Different researchers and educators believe that the ultimate goal of educational assessments is to provide diagnostic information. Thus, despite their essential contribution of summative assessments to educational measurement, another type of assessment is also needed. This diagnostic information provided in a timely fashion can be used, for example, to create specifically developed remedial instructional programs to address students' deficiencies. These assessments have been also referred to as *cognitive diagnostic assessments* (CDAs; de la Torre and Minchen 2014). The statistical models that are capable of supporting the CDAs are the CDMs. The focus of this dissertation, as mentioned earlier, is to continue with this innovative area of research.

Unlike traditional IRT models, which generally model continuous latent variables, the latent variables in CDMs are discrete, consisting either of dichotomous (e.g., mastery vs non-mastery), or polytomous

levels (e.g., "good performance", "fair performance", and "poor performance"). Through this section, we will illustrate the main characteristics of CDMs using simulated data: their multidimensional nature, their confirmatory nature, the complexity of their loading structure, and the type of latent predictor variables they contain (Rupp and Templin, 2008). The dataset used for illustration is included in the **GDINA** R package (Ma and de la Torre, 2017). It includes the responses of 1,000 respondents to 10 items measuring 3 attributes. This dataset was selected first because items were generated using some of the different CDMs that will be discussed in this Section 1.2; and, second, because its simplicity in terms of number of items and attributes.

As we said before, CDMs are inherently confirmatory. This is indicated by their loading structure, which is commonly known as Q-matrix (Tatsuoka, 1983). The Q-matrix is a mapping structure that indicates the skills required for successfully answering each individual item. In the CDMs literature there is a consistent notation that will be generally employed in this dissertation. Respondents (e.g., learners, patients, applicants) are indexed by $i = 1, \ldots, I$, assessment items are indexed by $j = 1, \ldots, J$, and attributes (e.g., borrowing numbers, a diagnostic criteria for pathological gambling, a professional competency) are indexed by $k = 1, \ldots, K$. Observed responses of respondent i to item j are denoted x_{ij}, while the latent class (i.e., profile vector) of a respondent is denoted $\boldsymbol{\alpha}_i$, such that α_{ik} indexes whether respondent i has mastered skill k ($\alpha_{ik} = 1$) or not ($\alpha_{ik} = 0$). A Q-matrix can be viewed as a cognitive design matrix that makes explicit the internal structure of a test. The Q-matrix used in this illustration is displayed in Table 1.1. The Q-matrix is a $J \times K$ matrix of zeros and ones, where the element on the jth row and kth column of the matrix, q_{jk} indicates whether skill k is involved in answering item j ($q_{jk} = 1$) or not ($q_{jk} = 0$). For simplicity, the description of the methods in this section assumes an optimal performance assessment. In the context of typical performance assessments, α_{ik} represents whether respondent i presents attribute k, and q_{jk} whether attribute k affects the probability of endorsing item j.

Table 1.1: Q-Matrix for the simulated dataset

Item	α_1	α_2	α_3
1	1	0	0
2	0	1	0
3	0	0	1
4	1	0	1
5	0	1	1
6	1	1	0
7	1	0	1
8	1	1	0
9	0	1	1
10	1	1	1

Note. $\alpha_{1...3}$: Attributes being measured by the test.

As can be seen from the Table 1.1, three items involved only one attribute, six items involved two attributes, and one item involved three attributes. Confirmatory Factor Analysis (CFA) models and IRT models usually have a simple structure, that is, each item loads only in one factor (for a detailed discussion, see McDonald 1999). Factors, as defined in these models, are generally broader dimensions (e.g., number ability). On the contrary, in the case of CDMs, factors, commonly referred to as *attributes*, are narrowly defined (e.g., fraction subtraction). Each item typically requires more than one attribute. This leads to a complex loading structure where each item is specified in relation to multiple attributes. This complex loading structure, in terms of multidimensional IRT, is known as within-item multidimensionality (Adams et al., 1997) and is reflected in the "1s" of the Q-matrix as it happens, for example, in the componential IRT models (Embretson, 1991; Fischer, 1997).

The following provides a graphical comparison of a few of the prototypical models that we have discussed. Figure 1.1 depicts three different psychometric models so that we could better understand the difference between simple structure and complex structure. Note that the horizontal lines for categorical variables reflect thresholds (i.e. the probability of a respondent possessing or mastering dichotomous attributes and probabilities of correct response for dichotomous observed responses). In these figures, these bars are located at arbitrary points to simplify the illustrations. Figure 1.1 A and Figure 1.1 B shows a three-dimensional CFA and IRT models with simple structures and contrast it with Figure 1.1 C, which shows a three-dimensional CDM with a complex loading structure (i.e. items 4, 5, 6, 7, 8, 9, and 10 load on several dimensions). In this way, CDMs could be understood as an extension of traditional multidimensional IRT and CFA models that is particularly suitable for modeling complex loading structures.

In short, CDMs are latent class models (Hagenaars and McCutcheon, 2002) that classify respondents into some latent classes according to similarity of their responses to test items. They are called restricted latent class models because the number of latent classes is restricted by the number of attributes involved in answering items of a test. With K attributes underlying performance on a given test, the respondents will be classified into 2^K latent classes (the number 2 indicates that there are two possible outcomes for each attribute: mastery or non-mastery). Latent classes are indexed by $l = 1, \ldots, 2^K$. In our CDM example, there are three attributes required to perform successfully on the items. Thus, test takers will be classified into $2^3 = 8$ latent classes. Table 1.2 shows the attribute class probabilities of the sample composed of 1,000 respondents classified into the eight possible latent classes using a CDM.

The main output of CDM for each respondent is a vector of estimates denoting in terms of posterior probability the state of mastery on each of the attributes. These probabilities are typically converted in dichotomous scores (i.e., mastery or non-mastery) by comparing them to a cut-off score (usually .5; de la Torre et al. 2010; Templin and Henson 2006) to define these attribute profiles. An example of the output

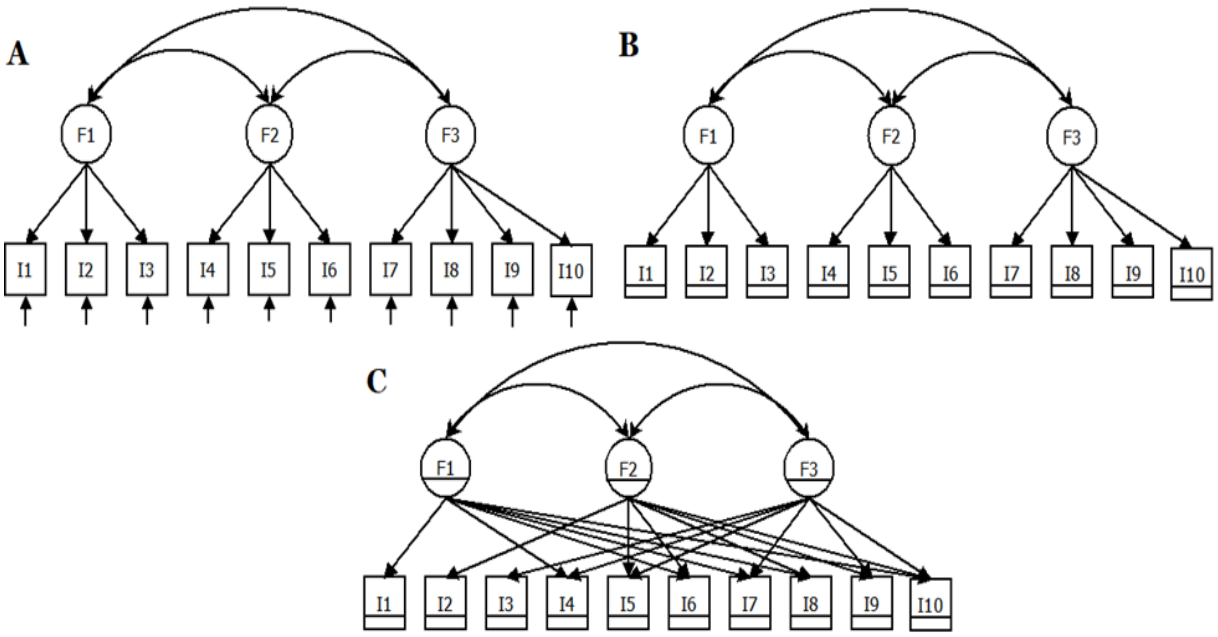

Figure 1.1: Representation of the different prototypical models. Model A = Three-dimensional CFA model with simple loading structure; Model B = Three-dimensional IRT model with simple loading structure; Model C = Three-dimensional CDM with complex loading structure. $F_{1...3}$: Different dimensions being measured by the 10-item test. This figure has been adapted from Rupp and Templin (2008).

for a respondent is depicted in Figure 1.2. This examinee will have probably mastered attributes 1 and 3, but have not mastered attribute 2. The degree of uncertainty of an estimate near .50 is high. Following this, different researchers suggest an *indifference* region between .40 and .60 where no classifications are stablished (e.g., Sorrel et al. 2016; Templin and Henson 2006).

In the field of education, researchers have proposed different theories about how students represent knowledge and develop competence in a subject domain (e.g. Mathematics). In this sense, the CDM approach fits in very well with the actual trends in cognitively diagnostic assessments in education (Leighton and Gierl, 2007; Nichols et al., 2012). In addition, despite the fact that few empirical studies have been published out of the educational context, CDMs can be applied to other contexts. In this regard,

Table 1.2: Latent class probabilities

Latent Class	Attribute profile	Posterior probability
1	000	.1274
2	100	.1084
3	010	.1129
4	001	.1193
5	110	.1310
6	101	.1466
7	011	.1391
8	111	.1153

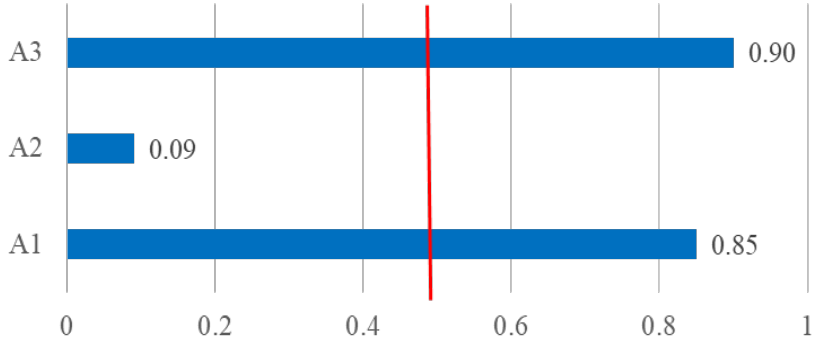

Figure 1.2: Output of the CDM analysis at the examinee level. An horizontal line is included at 0.50, which is the usual cut-off value to distinguish between mastery and non-mastery of each attribute. A1-A3 represents attributes 1 to 3.

there are two important studies that are good examples. First, the study of Templin and Henson (2006), which demonstrate how the hypothesized underlying factors contributing to pathological gambling can be measured with the deterministic input, noisy "or" gate (DINO; Templin and Henson 2006) model. Second, the study of García et al. (2014), who found that a different CDM, the generalized deterministic input, noisy "or" gate (G-DINA; de la Torre 2011) model, achieved an accurate fit to the responses of a situational judgement test (SJT) measuring six professional competencies based on the great eight model (Bartram, 2005). In the following, the variety of CDMs available is discussed (Subsection 1.2.2), specific insights on the CDMs used in this dissertation (Subsection 1.2.3) and a detailed review on the current empirical applications (Subsection 1.2.4) are provided. Finally, Subsection 1.2.5 and Subsection 1.2.6 review two of the areas of research where this dissertation is intended to contribute, namely model fit assessment and the use of CDMs for adaptive testing.

1.2.2 A Taxonomy of cognitive diagnosis models

Generally, CDMs can be grouped into three families as shown in Table 1.3, where some of the widely employed CDMs are included (different classifications can be found in, e.g., Rupp et al. 2010 and DiBello et al. 2007). Considering the manner in which the latent predictor variables are combined, CDMs can be divided into compensatory and non-compensatory models. In non-compensatory latent-variable models, a low value on one latent variable cannot be compensated by a high value on another latent variable whereas in compensatory latent-variable models a low value on one latent variable can be compensated by a high value on another latent variable. The deterministic input, noisy "or" gate (DINA; Haertel 1989) model is an example of noncompensatory CDM.

The DINA model has two parameters per item, namely the guessing (g_j) and the slip parameters (s_j). The guessing parameter indicates the probability that respondents who have not mastered at least one of the required attributes for item j correctly answer the item. The slip parameter indicates the

probability that respondents who have mastered all the required attributes for item j incorrectly answer the item. Some of the categories in Table 1.3 comprises several models. Each model has its own specific features. For example, the noisy input, deterministic and gate (NIDA; Junker and Sijtsma, 2001) model is a noncompensatory model just like the DINA model. However, in contrast to the DINA model, the guessing and slip parameters are specified at the attribute level (i.e., g_k and s_k).

Non-compensatory models are better aligned with cognitive theory in some cases in which it is strongly believed that the respondent must have mastered all the attributes within the item in order to get the item correct. This might be the case in math education where all the required skills are needed in order to solve a certain problem. The fraction subtraction data originally described and used by Tatsuoka (1990) and more recently by Tatsuoka (2002, 2005) and de la Torre (2011) can help us illustrating an item following noncompensatory model. This dataset consists of 12 fraction subtraction problems involving four attributes: (a) performing basic subtraction operations, (b) simplifying/reducing, (c) separating whole numbers from fractions, and (d) borrowing one from whole numbers to fractions. Item 10 included in Figure 1.3 requires attributes (a), (b), and (c), but not attribute (d). Probably, students mastering these three attributes will correctly answer this item. Consistently, a further study indicated that the DINA model has a good fit to this item (de la Torre and Lee, 2013). Indeed, the DINA model is very popular in applications in the area of education (e.g., Choi et al., 2015; Liu et al., 2013). It may be argued that when responding to items like the one in Figure 1.3, the lack of one attribute cannot be compensated by the mastery of a different one. In the context of this item, if the respondent does not know how to do any of the steps, most likely won't answer the item correctly.

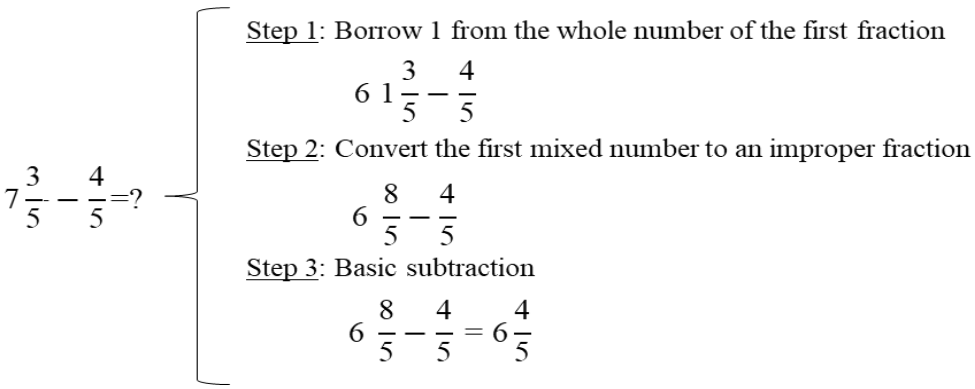

Figure 1.3: Sequential steps to solving item 10 of Tatsuoka (1990) dataset.

Different items might reflect different cognitive processes. Accordingly, different models were developed. Templin and Henson (2006)'s study is probably the first application of CDMs outside the educational context to a new area of application, namely clinical psychology. In this study, they introduced a compensatory CDM that is a disjunctive version of the DINA model, the deterministic inputs, noisy "or" gate (DINO) model. Templin and Henson (2006) argued that conjunctive models like the DINA model

Table 1.3: A taxonomy of CDM.

CDM Type	Data Type		Model	Model Name	Major References
Reduced	Dichotomous	Noncompensatory	RSM	Rule-space method	Tatsuoka (1983)
			DINA	Deterministic inputs, noisy "and" gate	Haertel (1989)
			NC-RUM	Non-compensatory reparametrized unified model	DiBello et al. (1995); Hartz (2002)
			NIDA	Noisy inputs, deterministics "and" gate	Junker and Sijtsma (2001)
			HO-DINA	Higher-order DINA	de la Torre and Douglas (2004)
			MS-DINA	Multiple-strategy DINA	Huo and de la Torre (2014)
		Compensatory	C-RUM	Compensatory RUM	Hartz (2002); Templin (2006)
			DINO	Deterministic inputs, noisy "or" gate	Templin and Henson (2006)
			NIDO	Noisy inputs, deterministic "or" gate	Templin (2006)
			A-CDM	Additive CDM	de la Torre (2011)
	Polytomous	Noncompensatory	RSM	Rule-space method	Tatsuoka (1983)
			NC-RUM		
			MC-DINA	Multiple-Choice DINA	de la Torre (2009a)
		Compensatory	C-RUM		
General	Dichotomous		GDM	General diagnostic method	von Davier (2005)
			LCDM	Loglinear cognitive diagnosis model	Henson et al. (2009)
			G-DINA	Generalized DINA model	de la Torre (2011)
	Polytomous		GDM		
			pLCDM	Polytomous LCDM	Hansen (2013)
			RS- and US-GDINA	Sequential G-DINA model	Ma and Torre (2016)
			GNDM	General nominal diagnosis model	Chen and Zhou (2017)

are not necessarily reasonable in the area of clinical psychology because a positive response to an item can happen for several reasons. To make this point clearer, Figure 1.4 includes one of the items used in their study and its Q-matrix specification. This item is taken from the Gambling Research Instrument (Feasel et al., 2004) which assesses the DSM–IV–TR Diagnostic Criteria for Pathological Gambling. According to the Q-matrix, this item measures two of the ten diagnostic criteria, namely (8) has commited illegal acts such as forgery, fraud, theft, or embezzlement to finance gambling, and (10) relies on others to provide money to retrieve a desperate financial situation caused by gambling. Disjunctive models like the DINO model allow for multiple strategies to solve a problem or multiple paths to give a positive response. Examinees meeting one or more of the criteria measured by the item will probably give a positive response. In the specific case of this item, a respondent will probably give a positive response if he or she has committed illegal acts, relies on others to provide him or her money, or both.

Item 22. Gambling has hurt my finantial situation

DSM-IV-TR Diagnostic Criteria	Q-matrix specification
(1) is preoccupied with gambling	0
(2) needs to gamble with increasing amounts of money in order to achieve the desired excitement	0
...	0
(8) has commited ilegal acts such as forgery, fraud, theft, or embezzlement to finance gambling	1
(9) has jeopardized or lost a significant relationship, job, or educational or career oportunity because of gambling	0
(10) relies on others to provide money to retrieve a desperate financial situation caused by gambling	1

Figure 1.4: One of the items employed by Templin and Henson (2006) when introducing the DINO model. Only a subset of the Q-matrix specification is presented.

With this in mind, compensatory models that assume a disjunctive process appear to be a reasonable option in the area of clinical psychology. However, it needs to be emphasized that reduced models like DINA and DINO make strong assumptions about the data and, because of that, their fit to the data should be carefully evaluated. This led de la Torre et al. (2017) to conduct model comparison analysis using data from another empirical application in the area of clinical psychology. These authors used 44 items from the Millon Clinical Multiaxial Inventory–III (MCMI-III; Millon et al. 2009) constituting the scales of anxiety, somatoform, thought disorder, and major depression. Perhaps surprisingly, they found that none of the items could be fitted using the DINO model, but 11 items could be fitted using the additive CDM (A-CDM; de la Torre 2011). The A-CDM considers the independent additive effects of the different disorders on the item endorsement. The rest of the items could not be considered as conjunctive,

disjunctive, or additive. A general CDM had to be then considered, that is, the generalized DINA model (G-DINA; de la Torre 2011).

General CDMs allow for all types of relationships within the same test. As will be discussed below, the G-DINA is a general CDM that subsumes several reduced CDMs. That means that different constraints can be imposed to the G-DINA model parameters, then obtaining the reduced models. This characteristic of the G-DINA model is crucial to the developments introduced in this dissertation and establishes the distinction between general and reduced CDMs. Given the wide range of CDMs (noted in Table 1.3), a critical concern is selecting the most appropriate model. Largely, model selection is a validation process given that the results of statistical models are meaningless when the model fit is poor. This dissertation deals with the evaluation of CDMs in a new area of application, examining item-level model fit indices, and improving CD-CAT results using model comparison indices. All of these developments are considered within the G-DINA model framework developed by de la Torre (2011). In the next section, we will provide more details about this framework.

1.2.3 The generalized DINA model framework

As shown by de la Torre (2011), many of the widely known CDMs can be represented via the G-DINA model, which is a generalization of the DINA model. As a general CDM, the G-DINA model allows to estimate a different model of each item on the same test. The G-DINA model describes the probability of success on item j in terms of the sum of the effects of the involved attributes and their interactions. This model partitions the latent classes into $2^{K_j^*}$ latent groups, where K_j^* is the number of required attributes for item j. Each latent group represents one reduced attribute vector $\boldsymbol{\alpha}_{lj}^*$ and has its own associated probability of success. This is the reason why this model is said to be saturated. The item response function is given by

$$P(X_j = 1|\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} \ldots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \tag{1.1}$$

where δ_0 is the intercept for item j (i.e., baseline probability), δ_{jk} is the main effect due to α_{lk}, $\delta_{jkk'}$ is the interaction effect due to α_{lk} and $\alpha_{lk'}$, and $\delta_{j12\ldots K_j^*}$ is the interaction effect due to $\alpha_1, \ldots, \alpha_{K_j^*}$. Thus, there are $2^{K_j^*}$ parameters to be estimated for item j.

In this dissertation, we will focus on three reduced models which are nested in the G-DINA model: the DINA model, the DINO model, and the A-CDM. These three models are good representatives of conjunctive (i.e., DINA), disjunctive (i.e., DINO), and additive (A-CDM) processes. Accordingly, results can be probably generalized to other similar CDMs (for more detailed information on the dissimilarities among different CDMs, see Ma et al., 2016). These three reduced CDMs are described below.

If several attributes are required to correctly answer the items, the DINA model is deduced from the G-DINA model by setting to zero all terms except for $\delta_j 0$ and $\delta_{j12...K_j^*}$. Therefore, the probability of success can be written as

$$P(X_j = 1|\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \delta_{j12...K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \tag{1.2}$$

That is, in the DINA model, except for the attribute vector $\boldsymbol{\alpha}_{lj}^* = \mathbf{1}$ (i.e., respondents mastering all the required attributes), the $2^{K_j^*-1}$ latent groups have identical probability of correctly answer the item j, that is, the baseline probability. As such, the DINA model has two parameters per item, commonly known as guessing (g_j) and slip (s_j) parameters.

The DINO model has also only two parameters per item, namely δ_{j0} and δ_{j1}. In order to derive the DINO model from the G-DINA model, the following constraint needs to be imposed:

$$\delta_{j1} = \delta_{jk} = -\delta_{jk'k''} = \ldots = (-1)^{K_j^*+1} \delta_{j12...K_j^*} \tag{1.3}$$

so that some lower-order terms will be cancelled by the corresponding higher-order terms.

When all the interaction terms are dropped, the G-DINA model reduces to A-CDM. The probability of a correct response for the A-CDM is given by

$$P(X_j = 1|\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk} \alpha_{lk} \tag{1.4}$$

This model indicates that mastering attribute α_{lk} increases the probability of success on item j by δ_{jk} independently of the contributions of the others attributes. The A-CDM has $K_j^* + 1$ parameters per item.

In order to facilitate the understanding of the differences among the models, Figure 1.5 depicts the parameter estimates for the G-DINA model for two items from the dataset used in this illustration. The vertical axis shows the point estimate for each parameter and the associated standard-error band (i.e., parameter value standard error); the red horizontal line indicates the value of 0 as a visual reference point. We can identify a likely candidate CDM for each item. For example, the pattern of parameter estimates for Item 5 shows that the main effects are essentially 0 ($\delta_1 = -0.03$ and $\delta_2 = -0.01$). This pattern is consistent with the DINA model where all the parameters except the baseline probability and the highest-order interaction are set to 0. In this item, $\delta_0 = 0.11$ and $\delta_{12} = 0.76$. Then, examinees lacking at least one of the required attributes will have a probability of success equal to 0.11, and examinees mastering the two required attributes will have a probability of success of $0.11 + 0.76 = 0.87$, approximately. The pattern of estimates for Item 8, on the other hand, shows that all the effects contribute to the probability of success. Each of the required attributes has a similar contribution ($\delta_1 = 0.15$ and $\delta_2 = 0.16$). The greatest increment in the success probability occurs when both attributes are mastered, as indicated by the

interaction effect ($\delta_{12} = 0.48$). This pattern is not consistent with any of the reduced models, and then the general model needs to be retained (i.e., the G-DINA model).

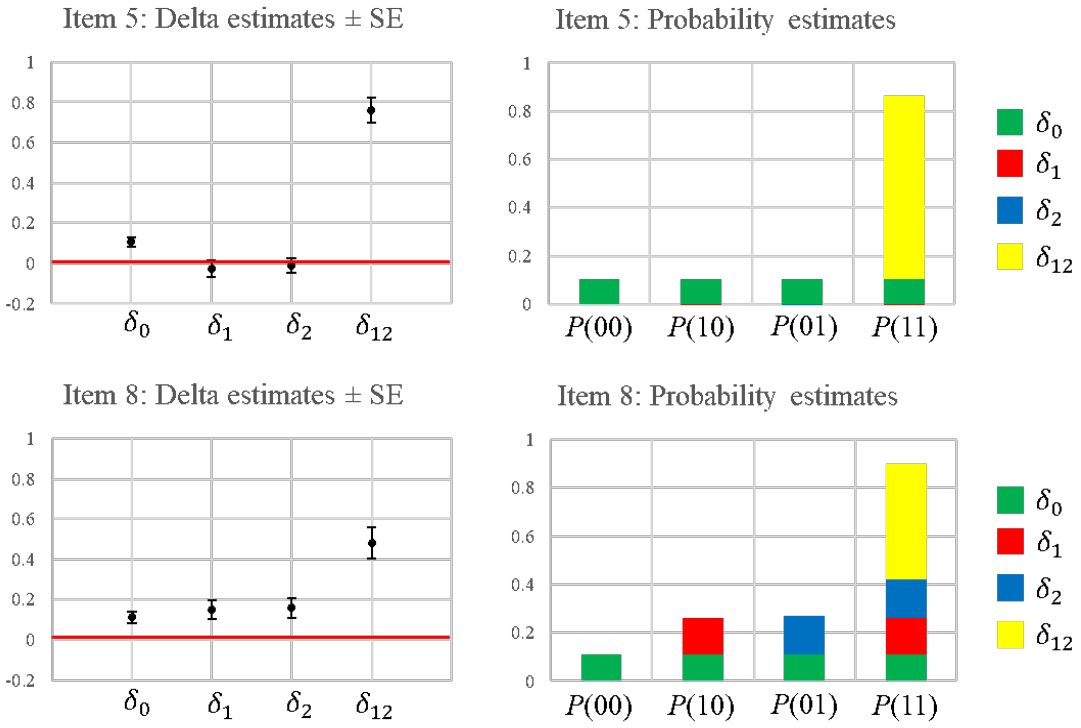

Figure 1.5: Parameter estimates for the G-DINA model for two example items and the derived probabilities of success for the different latent classes.

The marginal maximum likelihood method with Expectation-Maximization (MMLE/EM) algorithm is typically used for item parameter estimation under the G-DINA framework (e.g., **GDINA** R package; Ma and de la Torre 2017; **CDM** R package; Robitzsch et al. 2017). Specifically, for the G-DINA model:

$$P(X_j = 1|\boldsymbol{\alpha}_{lj}^*) = \frac{R_{lj}}{I_{lj}} \tag{1.5}$$

where R_{lj} is the expected number of examinees with attribute pattern $\boldsymbol{\alpha}_{lj}^*$ answering item j correctly and I_{lj} is the expected number of examinees with attribute pattern $\boldsymbol{\alpha}_{lj}^*$. For DINA or DINO model, R_{lj} and I_{lj} are collapsed for latent classes having the same probability of success. For A-CDM, some optimization techniques are adopted. Further details can be consulted in de la Torre (2009b, 2011). Different researchers have used Markov-Chain Monte Carlo in more complex situations, for example, when dealing with higher-order structures (e.g., de la Torre and Douglas 2004). However, this is not the case in most situations, and then the MMLE/EM algorithm that requires less time compared to the MCMC estimation is the one used more frequently. Person parameters are typically estimated using the maximum likelihood (ML), maximum a posteriori (MAP), or expected a posteriori (EAP) methods. These methods are described in detail in Huebner and Wang (2011).

1.2.4 Review on the current empirical applications

As previously noted, most of the CDM applications are in the area of education, where these models first emerged. The first application is probably the one by Tatsuoka (1990). This application in the domain of mixed-number subtraction has been already mentioned in this section. This same dataset has been used by many researchers in the context of CDM (e.g., Mislevy 1994; de la Torre 2011). CDMs have also been used with educational surveys for reading and mathematics assessments such as the National Assessment in Educational Progress (NAEP) (Xu and von Davier, 2006) and TIMSS (Choi et al., 2015; Lee et al., 2011); TOEFL (von Davier, 2005) and mock TOEFL tests (Liu et al., 2017); fraction arithmetic assessment (Bradshaw et al., 2014; Kuo et al., 2017); the evaluation of reading and listening comprehension (Baghaei and Ravand, 2015; Li et al., 2016; Ravand, 2016; Yi, 2017); and spatial reasoning in the context of student learning (Wang et al., 2018). Empirical CD-CAT applications have also taken place in the context of educational measurement, and focused on English language proficiency testing (Liu et al., 2013) and proportional reasoning assessment (Sorrel et al., 2018c).

Applications in other areas are scarce. Probably one of the most promising areas of application of CDM is the detection of psychological disorders. Different studies applied CDMs to detect pathological gambling using the DSM-III criteria (Templin and Henson, 2006) and anxiety, somatoform, thought disorder, and major depression using the MCMI-III (de la Torre et al., 2017). More recently, Tu et al. (2017) developed a questionnaire to measure the internet gaming disorder using CDMs based on DSM-V.

Other than that, the potential of CDMs has not been really exploited in other areas. A pioneering study by García et al. (2014) applied CDM to a SJT evaluating work competencies. This work inaugurates the application of CDMs to the area of Industrial-Organizational psychology. Later on, Sorrel et al. (2016) propose that CDMs can be the basis for a new framework to evaluate SJT, and specify the sequential steps in their application using data from an assessment of students' competencies. This publication is detailed in Chapter 2. There are new studies following this line of research, including Bley (2017) and Chen and Zhou (2017). Specifically, Chen and Zhou (2017) introduce a polytomous CDM that is suitable for SJT data. Bley (2017) describes an application of CDMs to measure intrapreneurship competence.

1.2.5 Model fit in cognitive diagnosis modeling

As in any statistical model, a prerequisite to study model results is to ensure that the model has an acceptable fit to the data. To this end, model fit should be examined at the test, person, and item levels. Chapter 3 and Chapter 4 describe two publications in the context of item-level model fit evaluation. A brief introduction to item and model fit evaluation in CDM is provided in the following, and readers are referred to Chen et al. (2013), Hu et al. (2016), Lei and Li (2016), and Sen and Bradshaw (2017) for a detailed discussion on performance of the fit indices. Person-level fit evaluation is out of the scope of this

dissertation. Summarizing, fit at the person level is typically evaluated using the generalized likelihood ratio test (Liu et al., 2009) or the hierarchy consistency index (Cui and Leighton, 2009). More recently, Cui and Li (2015) investigated the performance of the well-known statistic l_z and introduced the response conformity index.

Two types of model fit can be examined at the test and item levels, namely absolute fit and relative fit. Absolute fit consists of determining if a single model provides adequate fit to the data. Absolute fit indices explore whether some characteristics of the data can be well reproduced by the model. Different fit statistics have been used in the context of CDM, including: those based on the residuals between the observed and predicted proportion of correctly answered items and correlations and log-odds ratios of item pairs (Chen et al., 2013; Sinharay and Almond, 2007; Wang et al., 2015); item discrimination indices (de la Torre and Chiu, 2016; de la Torre, 2008); χ^2 and G statistics based on the observed and predicted item-pair responses (Rupp et al. 2010, pp. 266-270); and the mean absolute differences (Henson et al., 2009) between the observed and predicted item conditional probabilities of success and related root mean square error of approximation (Kunina-Habenicht et al., 2012). More recently, two different studies examined the performance of the M2 statistic in this context (Hansen et al., 2016; Liu et al., 2016). Additionally, Chapter 3 introduces $S - X^2$ Orlando and Thissen (2000, 2003) for the purposes of item-level absolute fit evaluation.

On the other hand, it is likely that more than one model can fit the data adequately. Relative fit indices (also called model comparison indices) are used to determine the best fitting model among a set of competing models. Relative model fit has generally been evaluated using conventional information criteria (Chen et al., 2013) such as the Akaike's information criterion (Akaike, 1974) and the Bayesian information criterion (Schwarz et al., 1978). Only a few studies have ever looked into item-level relative fit analysis (i.e., model comparison at the item level). First, de la Torre and Lee (2013) introduced the Wald (W) test for this purpose. This test requires the models to be nested. A model is said to be nested within another model if the former can be reduced to a special case of the latter by setting one or more constraints on its parameters. There are some relative indices specifically developed for evaluating if the inclusion of those constraints led to a significant loss of fit, including the above-mentioned W test. If it can be shown that both the reduced and the general model provide the same fit to the data, the reduced model should be retained according to Occam's razor principle. Later on, Ma et al. (2016) extended the simulation design of de la Torre and Lee (2013) by including additional factors. Chapter 3 introduces in the CDM context the other two classical tests used to compare different nested models (Buse, 1982), namely the likelihood ratio (LR) test and the Lagrange multiplier (LM) test. Chapter 4 introduces an approximation to the LR that is shown to be much more efficient, while maintaining the good statistical properties of the classical implementation of the LR test.

In summary, the model selection should involve both theory and empirical fit statistics. An important distinction is made between general and reduced CDMs. Given that reduced CDMs are constrained versions of the general CDMs, the absolute fit of the general CDMs will be always better. There are, however, a number of reasons to prefer reduced CDMs (de la Torre and Lee, 2013). First, general CDMs are more complex and thus require large sample sizes to be estimated accurately. Second, reduced models have parameters with a more straightforward interpretation. Third, the lack of parsimony, or overfitting, may result in a poor generalization performance of the results to new data. Finally, some studies showed that, compared to a general CDM, using the correct reduced CDM or the correct combination of reduced CDMs can lead to a higher accuracy, particularly when the sample size is small and the item quality is poor (de la Torre and Sorrel, 2017; Ma et al., 2016; Rojas et al., 2012). More details about this effect are provided in Chapter 5, which explore it under a CD-CAT context.

1.2.6 Cognitive diagnosis computerized adaptive testing

Researchers have been striving to develop the necessary methods to implement CDMs into an adaptive testing setting. This new area of research is referred to as *cognitive diagnosis computerized adaptive testing* (CD-CAT; Cheng 2009; Huebner 2010). The underlying item response models in CD-CAT are CDMs. To date, the DINA and G-DINA models have been used the current empirical applications and real-data CAT simulations (Liu et al., 2013; Sorrel et al., 2018c). Liu et al. (2013) employed a 352-item English language proficiency item bank, and Sorrel et al. (2018c) a 76-item proportional reasoning item bank. Items in CDM are typically complex (i.e., measure more than one attribute). This has enabled to use smaller item banks, compared to CATs based on traditional IRT. This is supported by simulation studies indicating that very promising accuracy results are obtained with short tests and item banks (Cheng, 2009; Kaplan et al., 2015).

A diagram including the components of an adaptive assessment listed by Weiss and Kingsbury (1984) is shown in Figure 1.6. Most of the research in CD-CAT has been devoted to the study of item selection rules (e.g., Cheng, 2009; Hsu et al., 2013; McGlohen and Chang, 2008; Xu et al., 2003; Yigit et al., 2018; Kaplan et al., 2015). One important difference between traditional CAT and CD-CAT applications is that the Fisher information statistic (Lehmann and Casella, 2006), which is widely used in the traditional formulation of CAT, cannot be applied in CD-CAT because attributes in CDM are discrete. Fortunately, the Kullback–Leibler information (Chang and Ying, 1996), which is an alternative information statistic, has been successfully applied in this context (Cheng, 2009; Kaplan et al., 2015). In addition, new item selection rules were developed or adapted, including the general discrimination index (Kaplan et al., 2015) and the Jensen-Shannon divergence index for continuous (Minchen and de la Torre, 2016) and polytomous data (Yigit et al., 2018).

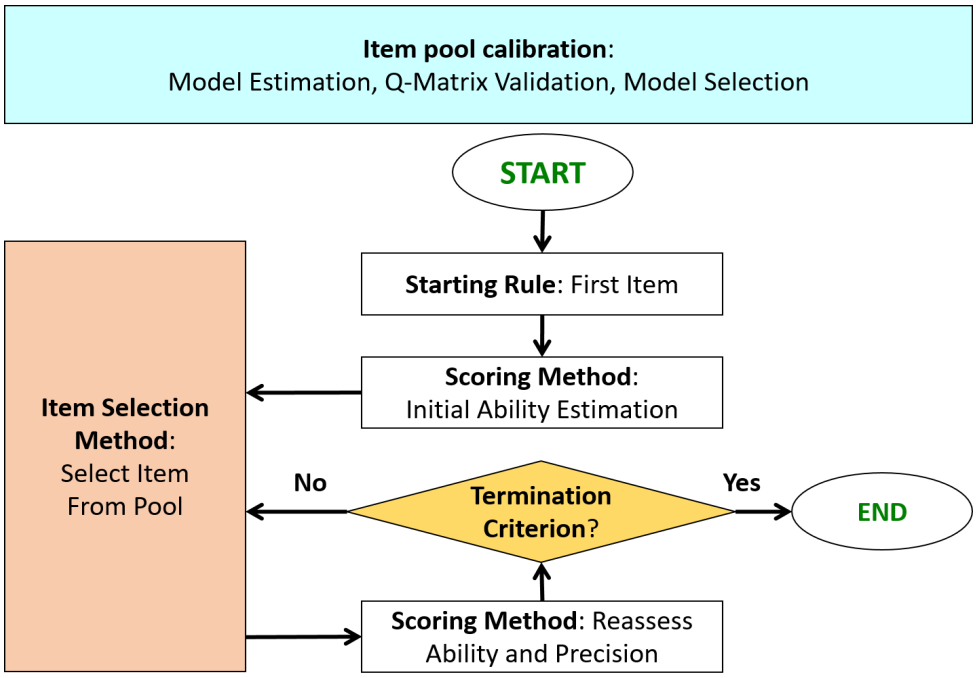

Figure 1.6: Adaptive testing flow diagram.

Regarding the termination criterion, both fixed and variable test length have been considered (Hsu et al., 2013; Kaplan et al., 2015). The criterion for the variable test length CD-CAT is the uncertainty in the respondent's classification. The CAT administration is typically terminated when the posterior probability that the respondent belonged to a given latent class is at least as large as a prespecified value. Four different values have been typically tested: 0.65, 0.75, 0.85, and 0.95. These values are then compared to the current latent class estimate, $\hat{\boldsymbol{\alpha}}_{\boldsymbol{ik}}$, computed using the MAP method. If the EAP estimator is considered, then the posterior probability of each individual attribute, $\hat{\alpha}_i$, is compared to the prespecified value and the CAT terminates when the condition is satisfied for all the individual attributes (Sorrel et al., 2018c).

Regarding the item bank calibration, the element that has received less attention is model selection. Due to its relative novelty, CD-CAT empirical applications are still scarce. A trend may be noted, however, towards the use of the same CDM for all the items in the item bank, either a reduced or a general one (Liu et al., 2013; Sorrel et al., 2018c). This will be the focus of Chapter 5, where it is examined whether better results can be obtained by evaluating item-level relative fit.

1.3 Goals of the current dissertation

The above describes, in short, the basis of the methodology introduced in this dissertation. Although the methods discussed in this dissertation are applicable to many fields, it is in the area of education where the methods have been more frequently applied. In addition, due to its relative novelty, some of the methods discussed should be further explored. All things considered, the main goal of this dissertation is to put

forward in the field of CDM. Specifically, there will be four studies focused on three different areas of research, namely the application of CDMs to other areas of knowledge, the study of model fit at the item level, and the item bank calibration in CD-CAT. Both real data and Monte Carlo methods are used in other to evaluate the current methods and the new proposals. The specific goals of the four studies are presented below.

1.3.1 Study 1: Application of cognitive diagnosis modeling to situational judgement tests data

The main goal of *Study 1* is to introduce CDMs as a new approach to evaluate SJT data. In this line, it is illustrated with an empirical example how CDMs can be used for evaluating the validity and reliability of SJT scores. The CDM approach is compared to the traditional approach based on CTT. This study follows a tutorial approach and it is intended to offer clear and easy-to-follow practical guidelines for researchers who work with SJT data.

Benefits of a psychometric model will not be effective unless it is ensured that the model fits the data. Accordingly, the rest of the studies focused on model fit evaluation.

1.3.2 Study 2: Inferential item fit evaluation in cognitive diagnosis modeling

The main goal of *Study 2* is to determine which of the item fit statistics performs best with CDM data. Specifically, four statistics are evaluated: $S - X^2$, the LR test, the W test, and the LM test. With the exception of the W test, it is the first time that the other three statistics are considered in the context of CDM. These statistics are compared in terms of Type I error and power using a Monte Carlo study. After a literature review, it is emphasized the need for the inclusion of item quality in Monte Carlo studies. This factor was overlooked in previous research, probably because item quality estimates in the area of education were generally high.

1.3.3 Study 3: Proposal of an approximation to the likelihood ratio test

The main goal of *Study 3* is to introduce the two-step LR (2LR) test as a new index for item-level model comparison. Result of *Study 2* indicated that the LR test was relatively more robust to the data factors than the other statistics. The current version of the LR test has the limitation to be very time consuming, given that it requires calibrating many different models and comparing them to the general model. The approximation that is introduced in this study only requires calibration of the more general model, so that this statistic may be easily applied in empirical research.

1.3.4 Study 4: Model selection in cognitive diagnosis computerized adaptive testing

The main goal of *Study 4* is to determine if better accuracy and item usage results can be obtained by using item-level model selection indices in CD-CAT. Specifically, the 2LR test is used to select the most appropriate model for each of the items in simulated item banks. Then, the performance of a CD-CAT based on a general model (i.e., G-DINA) and some reduced models (i.e., DINA, DINO, and A-CDM) is compared to that based on the combination of models selected with the two-step likelihood ratio test. For comparison purposes, the true item parameters are also considered, which allows obtaining an estimation of the upper limit for the classification accuracy.

Chapter 2

Validity and Reliability of Situational Judgement Test Scores: A New Approach Based on Cognitive Diagnosis Models

This study was published in the journal *Organizational Research Methods* (Sorrel, Olea, Abad, de la Torre, Aguado, & Lievens, 2016). The following pages include this publication. This article was first published on February 16, 2016. The issue was published on July 1, 2016. The article can be downloaded from https://doi.org/10.1177/1094428116630065. The journal impact factor and 5-year impact factor (2016) are 4.783 and 7.298, respectively.

Article

Organizational Research Methods
2016, Vol. 19(3) 506-532

DOI: 10.1177/1094428116630065
orm.sagepub.com

Validity and Reliability of Situational Judgement Test Scores: A New Approach Based on Cognitive Diagnosis Models

Miguel A. Sorrel[1], Julio Olea[1], Francisco J. Abad[1], Jimmy de la Torre[2], David Aguado[3] and Filip Lievens[4]

Abstract

Conventional methods for assessing the validity and reliability of situational judgment test (SJT) scores have proven to be inadequate. For example, factor analysis techniques typically lead to nonsensical solutions, and assumptions underlying Cronbach's alpha coefficient are violated due to the multidimensional nature of SJTs. In the current article, we describe how cognitive diagnosis models (CDMs) provide a new approach that not only overcomes these limitations but that also offers extra advantages for scoring and better understanding SJTs. The analysis of the Q-matrix specification, model fit, and model parameter estimates provide a greater wealth of information than traditional procedures do. Our proposal is illustrated using data taken from a 23-item SJT that presents situations about student-related issues. Results show that CDMs are useful tools for scoring tests, like SJTs, in which multiple knowledge, skills, abilities, and other characteristics are required to correctly answer the items. SJT classifications were reliable and significantly related to theoretically relevant variables. We conclude that CDM might help toward the exploration of the nature of the constructs underlying SJT, one of the principal challenges in SJT research.

Keywords

situational judgment tests, cognitive diagnosis models, validity, reliability

[1]Department of Social Psychology and Methodology, Universidad Autónoma de Madrid, Madrid, Spain
[2]Department of Educational Psychology, The State University of New Jersey, New Brunswick, NJ, USA
[3]Instituto de Ingeniería del Conocimiento, Universidad Autónoma de Madrid, Madrid, Spain
[4]Department of Personnel Management, Work and Organizational Psychology, Ghent University, Ghent, Belgium

Corresponding Author:
Miguel A. Sorrel, Department of Social Psychology and Methodology, Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, Madrid 28049, Spain.
Email: miguel.sorrel@uam.es

Situational judgment tests (SJTs) have become increasingly popular for personnel selection both in the United States and Europe (McDaniel, Morgenson, Finnegan, Campion, & Braverman, 2001; Whetzel & McDaniel, 2009). SJTs are designed to evaluate candidate judgments regarding situations encountered in the workplace (Weekley & Ployhart, 2006). Test takers are asked to evaluate each course of action either for the likelihood that they would perform the action or for the effectiveness of the action. SJTs are intended to evaluate different constructs (knowledge, skills, abilities, and other characteristics; KSAOs) related to job performance, which are different from those that are measured through cognitive ability tests or personality inventories. More specifically, a recent meta-analysis shows that SJTs intend to measure constructs that could be classified into four categories: knowledge and skills, applied social skills (e.g., leadership), basic personality tendencies (e.g., integrity), and heterogeneous composites (Christian, Edwards, & Bradley, 2010).

Despite their success, various validity and reliability issues related to SJTs have not been appropriately addressed (Christian et al., 2010; Ployhart & Weekley, 2006) because, as argued in the following, conventional methods for assessing the validity and reliability of SJT scores are based on classical test theory (CTT), which are inadequate in light of the multidimensional nature of SJT items. Therefore, this article explores the use of cognitive diagnosis models (CDMs) as a promising approach that not only overcomes these shortcomings but that also offers several advantages for scoring and better understanding SJTs.

The rest of the article is structured as follows. First, we briefly review existing validity and reliability evidence for SJT scores and in the process touch on the limitations of the existing approaches. The next section provides an introduction to CDMs. We then use an empirical example to illustrate how CDMs can be used for evaluating the validity and reliability of SJT scores and compare this approach with the traditional CTT approach. The last section discusses the advantages and the disadvantages of CDMs.

Review of SJT Literature on Reliability and Validity

Similar to any type of test, validation studies should also be conducted to provide relevant information for the interpretation and use of SJT scores. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 1999) specifies five ''sources of evidence that might be used in evaluating a proposed interpretation of test scores for particular purposes'' (p. 11). These sources of evidence are test content, consequences of testing, relations to other variables, internal structure, and response processes. In the following, we discuss to what extent these sources of evidence have been evaluated in the validation of SJT scores.

With regard to *evidence based on test content*, the norm in the development of SJTs is to recruit and train external ''subject matter experts'' (SMEs) to generate critical incidents. This information is used to develop the item stems, specify the extent to which these item situations represent the job domain, and establish the response alternatives and scoring key. Generally, once experts have made these decisions and judgments, the test is considered as more or less definitive. Furthermore, it is recognized that ''there is virtually no direct investigation of the relationships linking SJTs scores and test content'' (Schmitt & Chan, 2006, p. 147).

A more extensive strand of SJT studies focused on both intended and unintended *consequences of SJTs score interpretation and use*. Most of this research examined potential adverse impact of SJT scores, test taker perceptions toward various SJT formats, and the fake-ability of SJTs in comparison to traditional tests (for reviews, see Lievens, Peeters, & Schollaert, 2008; Whetzel & McDaniel, 2009).

Next, a voluminous stream of SJT validation studies scrutinized *evidence of the relation of test scores to a relevant criterion* (e.g., other constructs), their criterion-related validity with respect to

performance criteria, and their incremental validity over and above other more traditional measures (see the meta-analyses of McDaniel et al., 2001; McDaniel, Hartman, Whetzel, & Grubb, 2007). Generally, SJTs were found to have corrected validities in the mid .20s and exhibited incremental validity above and beyond traditional predictors, such as cognitive ability and personality (see also Clevenger, Pereira, Wiechmann, Schmitt, & Schmidt-Harvey, 2001; Weekley & Ployhart, 2005).

In comparison to this large body of research on the criterion-related validity of SJT scores, there is much less attention devoted to how constructs underlying SJTs are specified and examined (Arthur et al., 2014; Schmitt & Chan, 2006). The meta-analysis of Christian et al. (2010), for instance, reported that about one third of the papers published about SJTs did not indicate the construct measured, did not provide enough information about these constructs, or provided only the composite score. They concluded that ''test developers and researchers often give little attention to the constructs measured'' (Christian et al., 2010, p. 84). In other words, although SJTs seem to partly predict performance and enhance the criterion-related validity of traditional personality and cognitive ability test scores, the underlying reasons are not clear because little is known about the nature of the constructs measured by SJTs.

Therefore, it is widely acknowledged that more specific studies about the constructs underlying SJTs are needed (Ployhart & Ehrhart, 2003). In a recent review of personnel selection research, Ryan and Ployhart (2014) posited that among all current principal lines of research in SJTs, the exploration of the nature of the constructs is the most pressing one. Such construct-level information is pivotal because it offers several theoretical and applied advantages (Christian et al., 2010), namely, understanding deeper why some tests predict work performance better than others, comparing more clearly the effectiveness of different selection methods, reducing contamination by non–job-relevant constructs, and justifying the interpretation of the scores and their fair use.

To assess *the internal structure* of SJTs, one of the strategies in past research typically involved obtaining evidence via factor analytic techniques. However, the application of factor analytic techniques to SJT data almost always led to ''a plethora of factors that are difficult to interpret'' (Lievens et al., 2008, p. 430) as well as nonsensical factor structure solutions. Hence, it is recognized that ''there has been little success in understanding what SJTs really measure'' (Ployhart & Weekley, 2006, p. 346). Due to the uninterpretable factor analytic results, it has been posited that SJTs are ''construct heterogeneous at the item level, because one item, for example, may target several performance dimensions'' (Patterson et al., 2012, p. 853). Despite the multidimensional nature of SJTs, a single composite score is generally reported in SJT research and practice. All of these findings point to the necessity of alternative approaches for examining the internal structure (dimensionality) of SJTs and for obtaining ''new insights into understanding the constructs assessed by SJTs'' (Whetzel & McDaniel, 2009, p. 200).

Apart from lack of progress on how the internal structure of SJTs can be better understood, little is also known about the *response processes* that govern the ways in which individuals respond to SJT items. In fact, different possibilities exist regarding how individuals might respond to SJT items and solve them on the basis of their ability/skills. For instance, if a particular item includes several skills, are test takers required to master each of the skills to produce the most accurate answer (i.e., a noncompensatory model)? Or, could mastery of one of the skills compensate for the lack of mastery of the other skills (i.e., a compensatory model)? Unfortunately, these different possibilities in how individuals might respond to SJT items have not been examined with the appropriate psychometric models. As such, there exists a need for psychometric models that can provide information not only about the statistical quality of the items but also about the correspondence between the items and the targeted cognitive processes. In other words, psychometric models are needed to evaluate, among others, the appropriateness of compensatory and noncompensatory models to shed light on the item responding processes.

Finally, with respect to *reliability* of SJT scores, most studies have focused on internal consistency reliability (see review of Lievens et al., 2008). Generally, the internal consistency indices reported in the SJT literature are typically low. For example, a mean of .46 was obtained in some meta-analyses (e.g., Catano, Brochu, & Lamerson, 2012). These low internal consistency reliability values do not necessarily indicate poor precision of measurement. Rather, these results could reflect the fact that Cronbach's alpha is not appropriate for assessing the reliability of multidimensional tests such as SJTs because Cronbach's alpha requires that the construct domain be homogeneous (Schmidt & Hunter, 1996). In this context, homogeneity refers to unidimensionality (i.e., items measure a single latent construct). Given the heterogeneity of SJTs, even at the item level, researchers should look for other approaches for estimating reliability. Among other approaches, it has been proposed that test-retest reliability might be a particularly better measure for assessing the reliability of SJT scores (Lievens et al., 2008; Whetzel & McDaniel, 2009). However, "in most operational situations . . . it is impractical to obtain test-retest data" (Catano et al., 2012, p. 344). This underscores the needs to find other, more practicable approaches to estimate reliability of SJTs.

To recap, our review of research on the validity of SJT scores shows that prior research thus far has mainly focused on approaches to establishing validity evidence on the basis of test content, testing consequences, and relations to other variables. In contrast, there have been few successful attempts in providing evidence about the internal structure and response processes involved in solving SJT items. Moreover, our review of prior research highlighted the problems with using factor analytic techniques and Cronbach alpha for multidimensional tests such as SJTs. Our review also makes it clear that reliance on CTT has hampered further progress on these unexplored issues, which by nature are complex and may require more advanced psychometric models.

Thus, given these shortcomings in existing research on the validity and reliability of SJT scores, a new psychometric approach in examining the nature of constructs in SJTs is needed. Consistent with recommendations from a recent review on SJT research (Weekley, Hawkes, Guenole, & Ployhart, 2015, p. 301), we propose a specific set of latent trait measurement models, namely, cognitive diagnosis models, as an alternative psychometric approach to obtain evidence on the validity of SJT scores, assess their reliability, and score the different KSAOs that are theoretically measured by the SJT.

Cognitive Diagnosis Models: A Tutorial

In the past few years, there has been an increasing interest in psychometric models referred to as cognitive diagnosis models. CDMs are latent trait measurement models that explicitly allow for inferences about the underlying cognitive processes involved in responding to items and the manner in which these processes interact. In this sense, CDMs establish a link between cognitive psychology and statistical modeling. Earlier applications of CDMs are found in *cognitively diagnostic educational assessment* (Leighton & Gierl, 2007; Nichols, Chipman, & Brennan, 1995). The information that these models provide has been used for diagnosing students' strengths and weaknesses, thereby giving teachers information that can be used to design instruction and intervention.

CDMs emerged from different fields: theory of classification (restricted latent class models; Haertel, 1989), item response theory (linear logistic test model; Fischer, 1973), and mathematical psychology (knowledge space theory; Doignon & Falmagne, 1999). Based on these different approaches, CDMs have many labels (e.g., *cognitively diagnostic models*, Henson & Douglas, 2005, *cognitive psychometric models*, Rupp, 2007; *structured IRT models*, Rupp & Mislevy, 2007).

CDMs are multidimensional, categorical latent-trait models developed primarily for assessing examinee mastery and nonmastery of a set of skills (e.g., competencies, task, knowledge, and cognitive process). Unlike traditional item response theory (IRT) models, which generally involve continuous latent variables, CDMs involve latent variables that are binary (e.g., mastery vs.

Table 1. Initial Q-matrix.

	Attribute			
Item	1. Study Habits	2. Study Attitudes	3. Helping Others	4. Generalized Compliance
1	1	0	0	0
2	0	1	1	0
3	1	0	0	0
4	0	0	0	1
5	1	1	0	0
6	1	0	0	0
7	1	0	0	0
8	1	1	0	1
9	1	0	0	0
10	0	0	1	0
11	1	1	0	0
12	1	0	1	0
13	1	0	0	1
14	1	1	0	0
15	1	1	0	1
16	1	0	0	0
17	0	0	1	0
18	0	0	1	0
19	1	1	0	1
20	1	0	0	0
21	1	1	0	0
22	0	1	1	1
23	0	1	1	0

Note: 1 = the attribute is required to choose the most effective response option; 0 = the attribute is not required to choose the most effective response option.

nonmastery). In the CDM literature, these categorical latent variables have been generically referred to as *attributes*. The number of attributes is denoted by K, and the attribute profile of respondent i is denoted by $\boldsymbol{\alpha}_i = \{\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK}\}$, where $\alpha_{ik} = 1$ or 0 represents mastery or nonmastery of attribute k, respectively. CDMs are inherently confirmatory in nature as they involve a prespecified loading structure. The loading structure of a CDM, which is commonly known as Q-matrix (Tatsuoka, 1983), is a mapping structure that indicates the KSAOs required for successfully answering each individual item. A Q-matrix can be viewed as a cognitive design matrix that makes explicit the internal structure of a test. Table 1 shows the initial Q-matrix for the 23-item SJT that will be used as example in this article.

As can be seen from the table, for this test, $K = 4$ attributes are assumed to underlie the process of responding. Consider the first five items in Table 1: Items 1 and 3 require attribute 1 only; item 2 requires both attributes 2 and 3; item 4 requires attribute 4 only; and item 5 requires both attributes 1 and 2. Items 1 and 2 are shown in Figure 1. Item 1 measures study habits. Students who engage in regular acts of studying probably will answer this item correctly. Item 2 measures study attitudes and helping others. More likely than not, students who approve the broader goals of education (e.g., education should be within everyone's reach) and tend to help others will correctly answer this item.

Confirmatory factor analysis (CFA) models and IRT models usually have a *simple structure*, that is, each item loads only on one factor (for a detailed discussion, see McDonald, 1999). Factors as defined in these models are generally broader constructs (e.g., numerical ability). In contrast, in the case of CDMs, attributes are more narrowly defined (e.g., converting a whole number to a fraction).

ITEM 1: When studying for an exam, do you find that you reach best results when:

a. **you start planning and setting aside time in advance**
b. work in a clean environment, even if it means taking time away from studying
c. wait for inspirations before becoming involved in most important study tasks
d. wait until the last day or so to study, knowing that you have to get it done now

ITEM 2: Your professor announces in class that undergraduate students are needed to help run subjects for his upcoming study. While you would not receive any formal sort of extra credit, the professor would appreciate any volunteers. Given the following choices, which option would you choose?

a. Examine your schedule and offer to volunteer a couple hours a week when it is personally convenient.
b. **Examine your schedule and offer to volunteer as many hours as you can**.
c. Realize that you would have to give up some of your free time and choose not to volunteer.
d. Offer to run subjects only if you are paid.

Figure 1. Items 1 and 2 of the situational judgment test (Peeters & Lievens, 2005). Most appropriate answer is shown in bold.

In addition, each item typically requires more than one attribute. This leads to a *complex loading structure* where each item is specified in relation to multiple attributes. This complex loading structure, in terms of multidimensional IRT, is known as *within-item multidimensionality* (Adams, Wilson, & Wang, 1997) and is denoted by "1s" in the Q-matrix. As noted by Schmitt and Chan (2006), SJTs tend to be multidimensional, even at the item level. Thus, in SJTs it is necessary for items to load on more than one factor. CDMs could be understood as an extension of traditional multidimensional IRT and CFA models that are particularly suitable to this kind of construct and complex loading structure.

CDMs are also called *restricted* (i.e., confirmatory) latent class models because the number of latent classes is restricted by the number of attributes involved in answering items of a test. With K attributes underlying performance on a given test, the respondents will be classified into 2^K latent classes (the number 2 indicates that there are two possible outcomes for each attribute, as in, mastery or nonmastery). A generic latent class or attribute profile can be denoted by $\boldsymbol{\alpha}_l$, where the subscript index goes from $l = 1$ to 2^K. Thus, in the aforementioned example with four attributes required to perform successfully on the test items, respondents will be classified into $2^4 = 16$ latent classes. All CDMs are expressed by $P(X_j = 1|\boldsymbol{\alpha}_l)$, the conditional probability of success on item j given the latent class l. The main output of CDM for each test taker is an estimate of the attribute profile, which gives the probability that the ith respondent has mastered each of the attributes. These attribute profile estimates are obtained using the expected a posteriori (EAP) method.[1] This probability can be converted into dichotomous scores (i.e., mastery or nonmastery) by comparing them to a cut-off point (usually .50; de la Torre, Hong, & Deng, 2010; Templin & Henson, 2006). Other authors (e.g., Hartz, 2002; Jang, 2005) define an uncertainty region (e.g., between .40 and .60) within which no classifications are made, thus requiring stronger evidence before conclusions about the respondent's state of mastery with respect to a particular attribute can be drawn.

A general CDM, called the *generalized deterministic inputs, noisy "and" gate* (G-DINA) model, was proposed by de la Torre (2011). The G-DINA model describes the probability of success on item j in terms of the sum of the effects of involved attributes and their interactions. This model partitions

the latent classes into $2^{K_j^*}$ latent groups, where K_j^* is the number of attributes required for item j. For example, Item 2 in Figure 1 requires two of the four attributes. These two attributes lead to four latent groups: those who mastered both attributes, one of the attributes, or none of the attributes. Each latent group represents one reduced attribute vector α_{lj}^* and has an associated probability of success, written as

$$P(X_{ij} = 1|\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_{j-1}^*} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} \ \ldots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk'}$$

where δ_{j0} is the intercept for item j (i.e., the probability of a correct response to an item when none of the required attributes for the item has been mastered), δ_{jk} is the main effect due to α_k (i.e., the change in the probability of a correct response as a result of mastering a single attribute), $\delta_{jkk'}$ is the interaction effect due to α_k and $\alpha_{k'}$ (i.e., the change in the probability of a correct response due to the mastery of both attributes), and $\delta_{j12\ldots K_j^*}$ is the interaction effect due to $\alpha_1, \ldots, \alpha_{K_j^*}$ (i.e., the change in the probability of a correct response due to the mastery of all the required attributes).

The G-DINA model subsumes several commonly encountered CDMs. These include the DINO (*deterministic input, noisy "or" gate*; Templin & Henson, 2006) and DINA (*deterministic input, noisy "and" gate*; Haertel, 1989; Junker & Sijtsma, 2001) models. If several attributes are required for correctly answering the items, the DINA model can be obtained from the G-DINA model by setting to zero all terms except for δ_0 and $\delta_{j12\ldots K_j^*}$; in the case of DINO model, there are also only two parameters per item, namely δ_0 and δ_{jk}, with the important exception that δ_{jk} is constrained to be equal to $\delta_{jk} = -\delta_{jk'k''} = \cdots = (-1)^{K_j^*+1}\delta_{j12\ldots K_j^*}$ for $k = 1, \ldots, K_j^*$, $k' = 1, \ldots, K_j^* - 1$, and $k'' > k', \ldots, K_j^*$, so that some lower-order terms will be cancelled by the corresponding high-order terms. The DINA is a noncompensatory model that divides respondents in those who have mastered all measured attributes and those who are lacking at least one measured attribute, whereas the DINO is a compensatory model that divides respondents in those who master at least one measured attribute and those who are lacking all measured attributes. In this respect, the DINA model involves a conjunctive process, whereas the DINO model involves a disjunctive process. Figure 2 gives a graphical representation of an item requiring two attributes when it conforms to the DINA model, the DINO model, or the more general model (i.e., the G-DINA model).

The characteristics of CDMs discussed previously make CDM suitable for modeling the responses to a SJT. We identify four sequential steps in the application of CDMs to SJTs (see Figure 3). The first step is to develop the Q-matrix. It involves specifying the skills that are underlying performance on the SJT items and an initial Q-matrix. Next, one evaluates whether some of the original attribute specifications need to be changed on the basis of the analysis of empirical data. Once the final Q-matrix has been determined, the second step is the selection of an appropriate CDM on the basis of absolute and relative model fit. The third step consists of interpretation of the item and person parameter estimates of the selected model. Finally, the fourth step consists of searching for validity and reliability evidence of the person parameter estimates. We follow these steps in our empirical example in the following.

Assessment of SJTs Through Cognitive Diagnosis Models

This article presents a new approach to the assessment of SJTs, which aims to account for the multidimensional structure of tests. It has been shown in a prior study (García, Olea, & de la Torre, 2014) that CDMs could achieve an accurate fit to SJT data and the scores obtained could be properly interpreted. The present article substantially extends this initial work by highlighting CDMs'

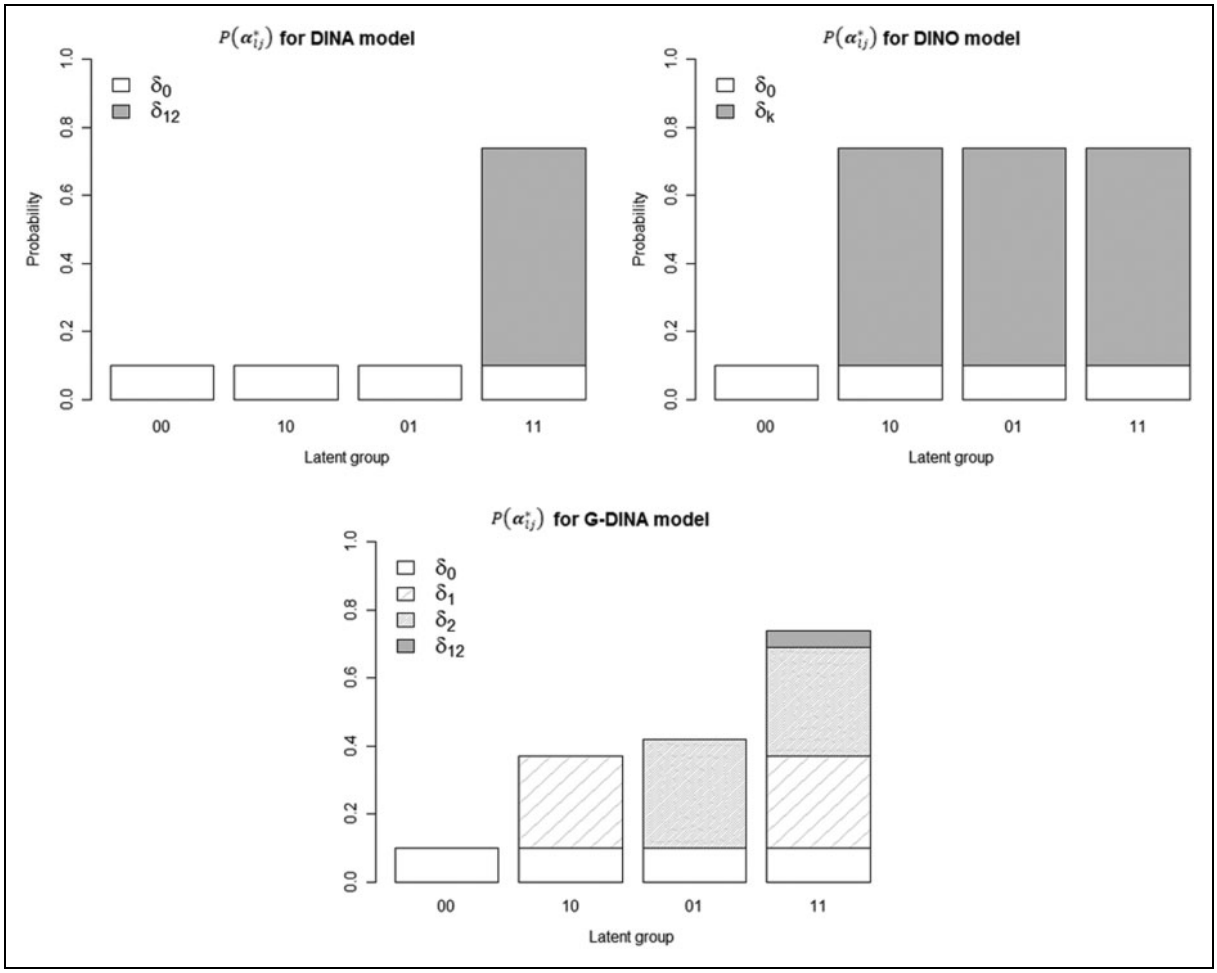

Figure 2. This figure depicts the probability of correctly answering an item requiring two attributes for *deterministic input, noisy "and" gate* (DINA), *deterministic input, noisy "or" gate* (DINO), and *generalized deterministic inputs, noisy "and" gate* (G-DINA) models. Model parameters are denoted by δ.

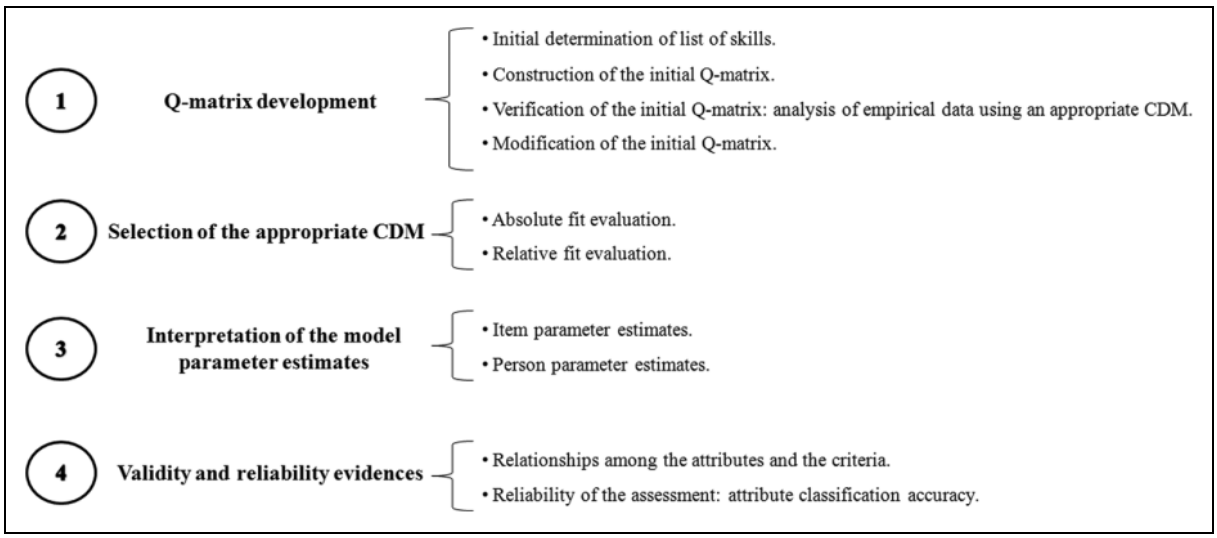

Figure 3. Sequential steps in the application of cognitive diagnosis models (CDMs).

usefulness in the context of reliability assessment and establishing the validity of SJTs. More specifically, this study is intended to address the following validity and reliability concerns:

1. What is the internal structure of the SJT? A CDM requires establishing a link between the items and the attributes through the Q-matrix specification. This task is typically conducted by domain experts. The recent empirical-based validation method proposed by de la Torre and Chiu (2015) then allows checking the Q-matrix generated by these experts. The Q-matrix specification and the model-fit results include information about the structural aspect, that is, how many attributes are involved at the test level, at the item level, and the relationships among them.
2. What is the general cognitive model that test takers engage in when responding to SJT items? The study of the absolute and relative fit of the different CDMs provides information about the general response processes required to solve the items. That is, we examine whether the sample of test takers engage in particular cognitive processes (e.g., conjunctive or disjunctive) when responding to the SJT.
3. Why are SJT scores good predictors of relevant theoretically relevant variables? As noted previously, SJT scores yield moderate criterion-related validity coefficients, and it is pivotal to better understand how and why SJT scores relate to the criteria and correlates. An explicit examination of the attributes measured by the SJT allows for this issue to be examined.
4. What is the reliability of the SJT assessment? As shown in the following, CDMs enable to address this question taking into account the heterogeneity of SJTs. We can use the calibrated model to generate simulate data, estimate the attribute profile for each test taker, and calculate the proportion of times that each test taker is classified correctly to the known attribute state (thus producing an estimate of attribute classification accuracy).

Demonstration Example

This section illustrates how CDMs can be applied to SJTs. The data for the present study were taken from the administration of an SJT composed of 23 items that present situations about various student-related issues (e.g., studying for exams and accomplishing assignments). This SJT was developed by Bess and Mullins (2002) and previously used by Peeters and Lievens (2005). By way of example, the first two SJT items are shown in Figure 1. As described in Peeters and Lievens, a total of 138 second-year psychology students from a large Belgian university participated in the study as a part of introductory courses about psychological testing and assessment. The sample was predominantly female (84%). The theoretically relevant variables (i.e., criteria and correlates) examined were grade point average (GPA, computed as the average of students' first- and second-year GPAs), student scores on the Advances Progressive Matrices (APM; Set II; Raven, Raven, & Court, 1998), and NEO Five-Factor Inventory (NEO-FFI; Costa & McCrae, 1992) self-report ratings (neuroticism, extroversion, openness to experience, agreeableness, and conscientiousness). Although the same data were used in Peeters and Lievens, CDM was not used in that study.

All the following analyses were carried out with the R (R Core Team, 2014) packages "CDM"[2] (Robitzsch, Kiefer, George, & Uenlue, 2015) (functions for cognitive diagnosis modeling) and "CTT" (Willse, 2014) (a function for classical test theory analysis). The code can be easily adapted to different data sets and can be requested by contacting the corresponding author.

Q-Matrix Development

As pointed out by Li and Suen (2013), when developing a new Q-matrix, it is common to adopt the following procedure (Buck et al., 1998):(a) Develop an initial list of skills, (b) construct an initial Q-

matrix, (c) analyze data using an appropriate CDM with the developed Q-matrix, and (d) modify the initial Q-matrix based on statistics for each skill along with the theoretical importance of the skill. We performed our analysis according to these steps.

Initial determination of list of skills. Given that the attributes are an essential part of the Q-matrix, it is important to use prior research, theory, and job analytic information for determining them. Other cognitive approaches such as think-aloud protocols have been also successfully employed to gather information about the possible cognitive processes (e.g., Li & Suen, 2013). Therefore, we relied on these information sources to come up with an initial list of attributes relevant to the SJT in our empirical example. In particular, our SJT consists of 23 items that present situations about various student-related issues. In the following, we outline the concepts that could underlie this specific SJT and how they might be linked to the theoretically relevant variables.

There is now relative consensus that performance comprises of both task and contextual performance (Motowidlo, Borman, & Schmit, 1997). Task performance involves behaviors that are directly relevant to core job functions, whereas contextual performance refers to behaviors to enhance the social and psychological climate in organizations. This theoretical distinction is made not only in the job performance domain but also in the academic performance domain (Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004).

Regarding dimensions underlying task performance in a student context, the meta-analysis of Credé and Kuncel (2008) revealed that study habits and study attitudes had incremental validity over standardized tests and previous grades in predicting academic performance (see also Aquino, 2011; Proctor, Prevatt, Adams, Reaser, & Petscher, 2006). Therefore, study habits and study attitudes were included in the initial list of attributes covered by the SJT in our example.

Regarding contextual performance, one of the main constructs covered is organizational citizen behavior (OCB; Bateman & Organ, 1983; Smith, Organ, & Near, 1983), which is defined by two major dimensions: (a) helping others and (b) generalized compliance (i.e., following rules and procedures; Organ, 1988). Importantly, such contextual activities are often similar across jobs and organizations (also known as transversal competences). Therefore, helping others and generalized compliance were also included in the initial list of attributes covered by the SJT in our example. Taking all of the aforementioned into account, an initial list of skills that was hypothesized to underlie this SJT was developed. Table 2 shows the four attributes (study habits, study attitudes, helping others, and generalized compliance) underlying performance on this SJT.

Next, we also put forward hypotheses related to the associations of these four attributes with theoretically relevant criteria and correlates. According to Motowidlo et al. (1997), variation in task performance is influenced by cognitive ability, whereas personality influences variation in contextual performance. Empirical findings have generally supported that personality factors predict contextual performance. In particular, three meta-analytic studies reported that conscientiousness, extraversion, neuroticism, and agreeableness are moderately correlated to cooperative contextual performance (Hough, 1992; Mount, Barrick, & Stewart, 1998; Organ & Ryan, 1995). LePine and Van Dyne (2001) found a similar pattern of results: Conscientiousness, extraversion, and agreeableness were more highly related to cooperative behavior than to task performance ($r = .17$ vs. $r = –.05$, $r = .14$ vs. $r = –.07$, and $r = .18$ vs. $r = .03$, respectively). The correlation between neuroticism and cooperative behavior, however, was not significantly higher than the correlation between neuroticism and task performance ($r = .05$ vs. $r = .09$). Openness was related to neither task performance nor cooperative behavior ($r = –.11$ and $r = –.07$, respectively). Although there exists less research on the generalized compliance dimension, Konovsky and Organ (1996) found that it was significantly related to conscientiousness ($r = .15$).

Table 2. Attribute Descriptions Based on Test Specifications.

Content Domain	Attribute	Definition	Typical Behavioral Patterns for People Mastering the Attribute in the Educational Environment
Task performance: (studies-related issues)	Study habits	Study habits refers to the pattern of behavior adopted by students in the pursuit of their studies that serves as the vehicle of learning. It is the degree to which the student engages in regular acts of studying that are characterized by appropriate studying routines occurring in an environment that is conducive to studying.	Reviews of material, study every day, take practice tests, efficiently organize his or her work, etc.
	Study attitudes	Study attitudes refers to a student's positive attitude toward the specific act of studying and the student's acceptance and approval of the broader goals of education.	Think education is relevant to their future, persist with enthusiasm or effort, have a good opinion of their teachers, etc.
Contextual performance: (transversal competencies)	Helping others	Helping others refers to voluntary actions that help another person with a problem. These helping behaviors can both be directed within or outside the organization.	Carry out volunteer actions that do not directly benefit them, share notes with their peers, help peers who are in troubles, etc.
	Generalized compliance	Generalized compliance refers to following rules and procedures, complying with organizational values and policies, conscientiousness, and meeting deadlines.	Stick with the existing timetable, be always punctual, do not defy the teacher, etc.

Concerning task performance, seven meta-analysis studies demonstrated consistent relationships between conscientiousness and task performance (the r coefficients vary from .20 to .31) across various occupational groups (Barrick & Mount, 1991; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Hurtz & Donovan, 2000; Salgado, 1997; Tett, Jackson, & Rothstein, 1991). Although it has been proposed that agreeableness may be an important predictor of task performance (Barrick & Mount, 1991), there is generally no evidence supporting this claim (Borman, White, & Dorsey, 1995; Hough et al., 1990; Hurtz & Donovan, 2000; Kamdar & Van Dyne, 2007; LePine & Van Dyne, 2001; Salgado, 1997).

Thus, given the aforementioned backdrop, we hypothesized that personality factors would be more highly related to the contextual performance dimensions of helping others and generalized compliance. Conversely, we hypothesized that cognitive ability and conscientiousness would be more highly related to task performance–related attributes such as study habits and study attitudes. In addition, we hypothesized that GPA would be more highly related to the studies-related attributes.

Construction of the initial Q-matrix. Four experts participated in an expert rating task. All of them were senior researchers with expertise in competency modeling and with extensive experience in teaching at the university level, and their native language was Spanish. The operational definitions of the four attributes were presented for their review and critique. The experts were asked to identify the

attributes needed for each item, thereby building the Q-matrix. The experts were also asked to specify the extent to which they were certain of their decisions. They employed the following system: 0 = it is certain that the attribute is not measured by the item, 1* = it is possible that the attribute is measured by the item, 1 = it is certain that the attribute is measured by the item. A Delphi process was used consisting of three rounds. In the first round, the experts were asked to identify the attributes needed for each item. In the second round, each Delphi participant was anonymously provided with the decisions of the other experts. This round provided an opportunity for participants to revise their judgments. Finally, in the third round, the four experts met in person and discussed in detail their opinions and settle the remaining differences. As done in Li and Suen (2013), we computed the Fleiss's kappa statistic (Fleiss, 1971) to evaluate the interrater reliability of the judgments made. We considered Landis and Koch's (1977) guidelines for interpreting kappa values, with values from .0 to .20 indicating a slight agreement, .21 to .40 a fair agreement, .41 to .60 a moderate agreement, .61 to .80 a substantial agreement, and .81 to 1 an almost perfect or perfect agreement. On the basis of the available evidence, we built the initial Q-matrix.

The experts' ratings across the three rounds are shown in Table 3. With regard to the first round, the Fleiss's kappa coefficients were .81 for helping others and generalized compliance and .53 for study habits indicating almost perfect and moderate agreements, respectively. However, the coefficient was only .17 for study attitudes. One possible reason for this is that this attribute is much more subjective than the other attributes, which made defining its behavioral outcomes more difficult. In the second round, when the experts were anonymously provided with the decisions made by the other experts, a high degree of agreement was achieved (the kappa coefficient for study attitudes increased up to .57). Finally, in the third round, a total agreement was achieved. The resulting attribute-item associations defined the initial Q-matrix (see Table 1). As can be seen, 11 items involved only one attribute, 8 items involved two attributes, and 4 items involved three attributes.

Verification of the initial Q-matrix: Analysis of empirical data using an appropriate CDM. There are many studies focused on the effect of Q-matrix misspecifications (e.g., de la Torre, 2008; Rupp & Templin, 2008a). In general, the results suggest that whenever a Q-matrix row is underspecified (i.e., a 1 is changed to a 0), the response probabilities for nonmasters of all measured attributes are overestimated (i.e., the items appear "easier"). In contrast, whenever a Q-matrix row is overspecified (i.e., a 0 is changed to a 1), we underestimate the response probabilities for masters of all measured attributes (i.e., the items appear "harder"). In addition, misspecifications in the Q-matrix may have important effects on the classification rates. Once the initial Q-matrix is specified, it is therefore important to verify its correctness. Otherwise, we cannot address any model misfit attributable to the Q-matrix.

To accomplish this, we used the test takers' responses to the SJT to empirically validate the Q-matrix following the general method of empirical Q-matrix validation recently proposed by de la Torre and Chiu (2015). This method is based on a discrimination index, which can be used in conjunction with the G-DINA model. Thus, the proposed index does not require making an assumption about which specific models are involved. The general discrimination index is defined as

$$\varsigma_j^2 = \sum_{c=1}^{2^{K_j^*}} w(\boldsymbol{\alpha}_{cj}^*)\,[P(\boldsymbol{\alpha}_{cj}^*) - \bar{P}_j]^2,$$

where $w\,(\boldsymbol{\alpha}_{cj}^*)$ is the probability of the reduced attribute pattern, $\boldsymbol{\alpha}_{cj}^*, P(\boldsymbol{\alpha}_{cj}^*)$ is the probability of success of the reduced attribute pattern $\boldsymbol{\alpha}_{cj}^*$, and $\bar{P}_j = \sum_{c=1}^{2^{K_j^*}} w(\boldsymbol{\alpha}_{cj}^*)P(\boldsymbol{\alpha}_{cj}^*)$ is the mean success probability. This discrimination index measures the extent to which an item can differentiate between the

Table 3. Expert Ratings for the Items of the Situational Judgment Test.

	Round				Round		
Item	1	2	3	Item	1	2	3
1	**1**	**1**	**1**	14	1^{a}, 2^{a}	1, 2	1, 2
2	1^{c}, 2^{a}, $\mathbf{3}^{*}$	2, 3	2, 3	15	**1**, 2^{c}, **4**	**1**, 2^{b}, **4**	1, 2, 4
3	**1**	**1**	**1**	16	**1**, 2^{c}	1	1
4	1^{c}, **4**	4	4	17	1^{c}, 2^{b}, $3^{a,*}$	1^{c}, 2^{b}, **3**	3
5	**1**, 2^{a}	1, 2	1, 2	18	3	3	3
6	**1**	**1**	**1**	19	$\mathbf{1}^{*}$, 2^{b}, **4**	**1**, 2^{b}, **4**	1, 2, 4
7	$\mathbf{1}^{*}$	**1**	**1**	20	$\mathbf{1}^{*}$, 2^{c}	**1**, 2^{c}	1
8	1^{b}, 2^{a*}, **4**	1^{a}, **2**, **4**	1, 2, 4	21	$1^{a,*}$, 2^{a}	**1**, 2^{a}	1, 2
9	$\mathbf{1}^{*}$	**1**	**1**	22	$1^{b,*}$, 2^{c}, 3^{b}, 4^{b*}	1^{b}, 2^{b}, 3^{b}, 4^{a}	2, 3, 4
10	$2^{c,*}$, **3**	3	3	23	2^{b}, $\mathbf{3}^{*}$, 4^{c*}	2^{a}, **3**	2, 3
11	$1^{a,*}$, 2^{a}	1, 2	1, 2				
12	1^{c}, 2^{c*}, 3^{a}, 4^{c}	1^{c}, 2^{c}, **3**, 4^{c}	1,3				
13	**1**, 2^{c}, $\mathbf{4}^{*}$	1, 4	1, 4				

Note: Attributes in bold were considered necessary by the four experts. Attributes: 1 = study habits; 2 = study attitudes; 3 = helping others; 4 = generalized compliance.
[a]Three experts considered the attribute necessary. [b]Two experts considered the attribute necessary. [c]One expert considered the attribute necessary.
[*]At least one expert expressed uncertainty about the necessity of the attribute.

different reduced attributed vectors based on their success probabilities and is minimum (i.e., equal to zero) when $P(\boldsymbol{\alpha}_1^* j) = P(\boldsymbol{\alpha}_2^* j) = \ldots = P\left(\boldsymbol{\alpha}^*_{2^{K_j^*} j}\right) = \bar{P}$. The maximum value of ς^2 for item j (i.e., ς^2_{jmax}) is obtained when all attributes are specified (de la Torre & Chiu, 2015). In addition, de la Torre and Chiu (2015) define the proportion of variance accounted for (PVAF) by a particular q-vector relative to this maximum as $\varsigma^2/\varsigma^2_{jmax}$.

Modification of the initial Q-matrix. As de la Torre and Chiu (2015) acknowledged, in many applied situations, Q-matrix recommendations based on the empirical validation procedure method can differ, sometimes markedly, from the Q-matrix based on expert opinions. In our case, changes suggested by the empirical validation were implemented if the following criteria were fulfilled: (a) gains in terms of the ς^2_j (i.e., ΔPVAF) were considered substantial (i.e., at least .30) and (b) changes made theoretical sense. To explore whether the changes suggested had theoretical basis, we took into consideration the ratings across the three rounds of the expert task (see Table 3). Note that experts were allowed to express uncertainty about their ratings (noted with * in Table 3). At this step, a suggested change was determined to have theoretical basis when at least one expert identified with certainty that the attribute as necessary/unnecessary. Finally, for the changes that met the criteria, we assessed the model fits with the Akaike Information Criterion (AIC; Akaike, 1974) to determine the final Q-matrix.

Although many of the suggested changes led to an improvement in the item discrimination, only Items 2 and 17 were found to also have some theoretical basis. For example, Item 2 in Figure 1 originally required attributes 2 and 3. As shown in Table 4, the suggested attribute specification prescribed all the attributes with ΔPVAF = .71. However, the experts recommended only attribute 1, but not attribute 4, with certainty (see Table 3, Round 1). This change has an associate ΔPVAF = .60. The same was true for item 17. To determine which of the suggested changes with theoretical basis to implement, we compared the model fit for four Q-matrix specifications, namely, the initial

Table 4. Largest $\hat{\varsigma}^2$ and PVAF of Item 2 for Different Numbers of Attribute Specifications.

Item	Attribute Specification	$\hat{\varsigma}^2$	PVAF
2	1000	0.05	.70
	0110[a]	0.02	.29
	1110	0.06	.86
	1111[b]	0.07	1.00

Note: $\hat{\varsigma}^2$ = general discrimination index; PVAF = proportion of variance accounted by the q-vector relative to the ς^2_{jmax}.
[a]Original. [b]Suggested.

Q-matrix, a change in Item 2 only, a change in Item 17 only, and changes in both Items 2 and 17. Based on the AIC, the best results were obtained for changing only the specification for Item 2. Therefore, we modified only the attribute specification for Item 2.

Selection of the Appropriate CDM

Each of the CDMs described in the introduction section specify the relationships among the postulated attributes in a different way. Whereas the DINA and DINO are conjunctive and disjunctive models, respectively, the G-DINA model is a general model that allows for both types of relationships within the same test. To select the most appropriate CDM for the test, one can assess the absolute and relative fit of each model. Considering that the DINA and DINO models are nested in the G-DINA model (de la Torre, 2011), one can employ the likelihood ratio (LR) test to evaluate their *relative* fit. The DINA and DINO models will always have a lower log-likelihood given that they are specific cases of the G-DINA model, but it is necessary to test whether the observed difference in model fit is statistically significant. The LR test does this by comparing the log-likelihoods of the models. This statistic is widely employed in other statistical models (e.g., structural equation models) for comparing nested models. It is assumed to be asymptotically χ^2 distributed with degrees of freedom equal to the difference between the numbers of parameters of the general and the reduced models. If the LR is significantly different from 0, the general model fits the data significantly better than the reduced model. Regarding *absolute* fit, we evaluated how well each proposed model reproduces the observed data. This is typically done by assessing indices based on residual analysis. We evaluated item fit statistics on the basis of the standardized residuals between the observed and predicted Fisher-transformed correlations of item pairs (Chen, de la Torre, & Zhang, 2013). To evaluate the absolute fit, Chen et al. (2013) proposed examining the z-score of the maximum absolute residual. If the evaluated model fits the data, this statistic should not be significantly different from zero. This approach is analogous to the inspection of the residual correlation matrix in structural equation modeling.

Table 5 shows the indices calculated for test fit and item fit for the G-DINA, DINA, and DINO models. The two χ^2 tests, each one with 44 degrees of freedom, corresponding to the likelihood ratio tests resulting from comparing the G-DINA model with the DINA (LR = 85.06) and DINO (LR = 82.55) models, were both significant ($p < .05$). These results indicate that the more parsimonious models led to a significant loss of fit. Absolute item fit statistics also indicated that the G-DINA model had better fit than the reduced models. When the G-DINA is fitted to the data, the z-score of the maximum absolute Fisher-transformed was not significant at α-level of .05 after applying the Holm-Bonferroni correction (Holm, 1979). Based on the previous information, the DINO and DINA model were discarded, and the G-DINA model was further examined for its adequacy to model the SJT data.

Table 5. Model Fit Indices for Different Cognitive Diagnosis Models.

			LR Test			Absolute Item Fit Statistics		
Model	loglike	Npars	LR	*df*	*p* Value	abs(fcor)	z-Score	*p* Value
G-DINA	−1,822.15	101				.28	3.28	.13
DINA	−1,864.68	57	85.06[a]	44	<.001	.32	3.75	.02
DINO	−1,863.43	57	82.55[b]	44	<.001	.32	3.71	.03

Note: loglike = log likelihood; Npars = number of model parameters; LR = likelihood ratio; abs(fcor) = maximum absolute Fisher-transformed correlation; DINA = *deterministic input, noisy "and" gate*; DINO = *deterministic input, noisy "or" gate*; G-DINA = *generalized deterministic inputs, noisy "and" gate*.
[a]G-DINA versus DINA. [b]G-DINA versus DINO.

Interpretation of Model Parameter Estimates

Item parameter estimates. In the next step, we described the items using both CTT and CDM indices. Regarding CTT indices, we used the proportion correct or item difficulty (P_j) and corrected point-biserial correlation (r_{cpb}). Based on the item parameter estimates for the selected CDM (G-DINA), $\hat{\varsigma}^2$ was computed. We also examined the difference between the probabilities of success for individuals who mastered none (i.e., $P(\mathbf{0}_j^*)$) and all of the attributes required (i.e., $P(\mathbf{1}_j^*)$). For example, if item j measures $K_j^* = 2$ attributes, this difference is computed as $P(11) - P(00)$ Unlike $\hat{\varsigma}_j^2$, this difference can be negative.

Table 6 presents the estimates of P_j, r_{pbc}, G-DINA parameters,$P(1_j^*) - P(0_j^*)$, and $\hat{\varsigma}_j^2$. In general, for the G-DINA model, good items are those that have small baseline probability (i.e., $P(0_j^*)$) and the probability of getting a correct response increases as the number of mastered attributes increases. For example, in the case of Item 5, the probability that respondent i with latent class α_i will correctly answer the item, an indicator for attributes 1 and 2, can be written as follows:

$$\begin{aligned} P(X_{i5} = 1|\boldsymbol{\alpha}_i) &= \delta_{50} + \delta_{51}\alpha_{i1} + \delta_{52}\alpha_{i2} + \delta_{512}\alpha_{i1}\alpha_{i2} \\ &= .62 + .07\alpha_{i1} + .27\alpha_{i2} + .04\alpha_{i1}\alpha_{i2} \end{aligned}$$

Thus, the baseline probability is rather high ($\delta_{50} = P(00) = .62$). The increment in the probability of correctly answering the item as a result of the presence of α_1 is small ($\delta_{51} = P(10) - P(00) = .69 - .62 = .07$), whereas mastering α_2 increases the probability of correctly answering the item up to .89 ($P(01) = \delta_{50} + \delta_{52} = .62 + .27 = .89$). The probability of success for respondents mastering both attributes is approximately 1 ($P(11) = \delta_{50} + \delta_{51} + \delta_{52} + \delta_{512} = .62 + .07 + .27 + .04 = 1$). The interaction effect due to the presence of both attributes is low ($\delta_{512} = P(11) - P(00) - P(10) - P(01) = 1 - .62 - .07 - .27 = .04$).

As can be seen from Table 6, some of the items with the lowest $\hat{\varsigma}^2$ had some of the highest $P(\mathbf{0}_j^*)$. For example, Item 13 was one of the least informative because nonmasters of the required attributes (1 and 4) have a substantial chance of guessing the correct answer, $P(00) = .75$. Indeed, it was found that a high percentage of the respondents answered the item correctly ($P_{13} = .91$).

To further explore the relationships between the G-DINA and CTT indices, the correlation between these indices was computed (see Table 7). We found a high significantly positive correlation between P_j and $P(\mathbf{0}_j^*)$ and $P(\mathbf{1}_j^*)$; the CTT discrimination index, r_{cpb}, was highly correlated with $P(\mathbf{1}_j^*) - P(\mathbf{0}_j^*)$ and moderately correlated to $\hat{\varsigma}_j^2$ and $P(\mathbf{0}_j^*)$. The two item discrimination indices in CDM were highly correlated.

Table 6. Classical Test Theory Indices and G-DINA Model Item Parameter Estimates.

			$P(\alpha_{lj}^*)$								Item Discrimination	
Item	P_j	r_{cpb}	P(0) P(00) P(000)	P(1) P(10) P(100)	P(01) P(010)	P(11) P(001)	P(110)	P(101)	P(011)	P(111)	$P(\mathbf{l}_j^*) - P(\mathbf{0}_j^*)$	$\hat{\varsigma}_j^2$
1	.71	.31	.42	.90							.48	.04
2	.35	.31	.00	.35	.36	.00	.02	.73	1	.54	.54	.06
3	.36	.28	.15	.49							.35	.02
4	.64	.21	.41	.76							.34	.00
5	.86	.36	.62	.69	.89	1					.38	.02
6	.52	.11	.43	.59							.16	.01
7	.60	.08	.62	.59							−.04	.00
8	.84	.07	.66	.62	1	1	1	1	.77	.83	.17	.01
9	.54	.14	.34	.67							.33	.04
10	.59	.25	.41	.76							.36	.00
11	.65	.32	.32	.68	.91	.75					.44	.01
12	.63	−.05	.78	.64	.00	.57					−.21	.01
13	.91	.16	.75	.95	1	.90					.16	.00
14	.43	−.01	.45	.00	.28	.54					.09	.01
15	.54	.29	.00	1	.00	.61	.85	.00	.67	.66	.66	.07
16	.69	.44	.30	.95							.65	.08
17	.37	.11	.24	.49							.25	.00
18	.49	.05	.45	.54							.09	.00
19	.27	.06	.38	.00	.00	.00	.38	.00	.00	.43	.05	.03
20	.40	.03	.32	.45							.12	.00
21	.55	.26	.17	1	1	.54					.37	.00
22	.72	.12	.58	.00	1	.57	.00	.89	.88	1	.42	.08
23	.12	.22	.00	.10	.00	.22					.22	.00

Note: $P(\alpha_{lj}^*)$ = probability of correctly answer the item for each latent group; P_j = item difficulty; r_{cpb} = corrected point biserial correlation; $\hat{\varsigma}_j^2$ = general discrimination index; G-DINA = *generalized deterministic inputs, noisy "and" gate*.

Table 7. Relationships Between Classical Test Theory Indices and G-DINA Item Parameter Estimates.

	P_j	r_{cpb}	$P(\mathbf{0}_j^*)$	$P(\mathbf{l}_j^*)$	$P(\mathbf{l}_j^*) - P(\mathbf{0}_j^*)$	$\hat{\varsigma}_j^2$
P_j	1					
r_{cpb}	.18	1				
$P(\mathbf{0}_j^*)$	.70**	−.46*	1			
$P(\mathbf{l}_j^*)$	.91**	.40	.51*	1		
$P(\mathbf{l}_j^*) - P(0_j^*)$	.13	.87**	−.57**	.42*	1	
$\hat{\varsigma}_j^2$	.11	.42*	−.25	.40	.65**	1

Note: P_j = item difficulty; r_{cpb} = corrected point biserial correlation; $\hat{\varsigma}_j^2$ general discrimination index; G-DINA = *generalized deterministic inputs, noisy "and" gate*.
*$p < .05$. **$p < .01$.

Person parameter estimates. Table 8 shows the attribute class probabilities and the class expected frequency in the sample of 138 respondents. The second column shows the possible attribute profiles for all the 16 latent classes. As the third column shows, the attribute profile of $\alpha_{16} = \{1111\}$ had the highest class probability of about .32. That is, approximately 32% of the respondents (as shown in the fourth column, 44 respondents) were classified as belonging to this latent class and therefore

Table 8. Estimated Occurrence Probabilities and Expected Frequency of the Latent Classes.

Latent Class	Attribute Profile	Class Probability	Class Expected Frequency
1	0000	.12	17.07
2	1000	.00	0.00
3	0100	.01	0.86
4	1100	.00	0.00
5	0010	.02	2.73
6	1010	.05	6.46
7	0110	.02	2.99
8	1110	.10	14.48
9	0001	.12	16.69
10	1001	.04	5.55
11	0101	.11	15.14
12	1101	.07	10.31
13	0011	.00	0.00
14	1011	0.01	1.11
15	0111	0.00	0.00
16	1111	0.32	44.61

were expected to master all of the four attributes. After applying the cut-off points (i.e., >.60 for mastery and <.40 for nonmastery), the percentage of examinees who did not receive a classification was 1%, 4%, 7%, and 2% for attributes 1, 2, 3, and 4, respectively.

Figure 4 depicts an example of how CDMs allow for a finer-grained analysis of the test takers' strengths and weaknesses. Test takers with the response pattern A correctly answered 9 items correctly. If we look at the Q matrix depicted in Table 1, we notice that these test takers correctly answer 4 out of the 6 items measuring generalized compliance (attribute 4). Thus, we estimate that they have a high probability (91%) of mastering this attribute. On this basis, these test takers are classified as masters of generalized compliance. Test takers with the response pattern B correctly answered 14 items correctly. We estimate that they have a high probability of mastering attributes 1, 2, and 4 (76%, 76%, and 93%, respectively). Note that despite the fact that these test takers fail at 6 out of the 10 items measuring study habits (attribute 2), some of the items that they correctly answered are highly discriminating (e.g., Items 5, 11, and 22). This explains why these test takers were estimated to have a high probability of mastering the attribute. The most uncertain estimate of an attribute mastery probability is at .50. For this reason, we recommend employing the discussed cut-off points (i.e., .40 and .60). Thus, no classification is made for helping others (attribute 3) for test takers with the response pattern B.

Validity and Reliability Evidences

Relationships among attributes and criterion/correlates. Once the person parameter estimates were estimated (i.e., the expected probability of mastering each attribute), we computed the correlations among the attribute scores, the SJT sum score, and the criterion/correlates. To eliminate the floor and ceiling effects inherent in the attribute probabilities, we used the logit transformation. As shown in Table 9, study habits (attribute 1) was highly correlated with GPA ($r = .35$) and conscientiousness ($r = .53$), and these correlation coefficients were somewhat higher than those estimates for the SJT sum score (.30 and .46, respectively). Thus, most of the predictive power of the SJT scores is due to this single attribute. Conversely, as we hypothesized, helping others (attribute 3) was generally related to the personality measures. The pattern of correlations is similar to the one obtained for the SJT sum score. Study habits and study attitudes (attributes 1 and 2) were also related to some of

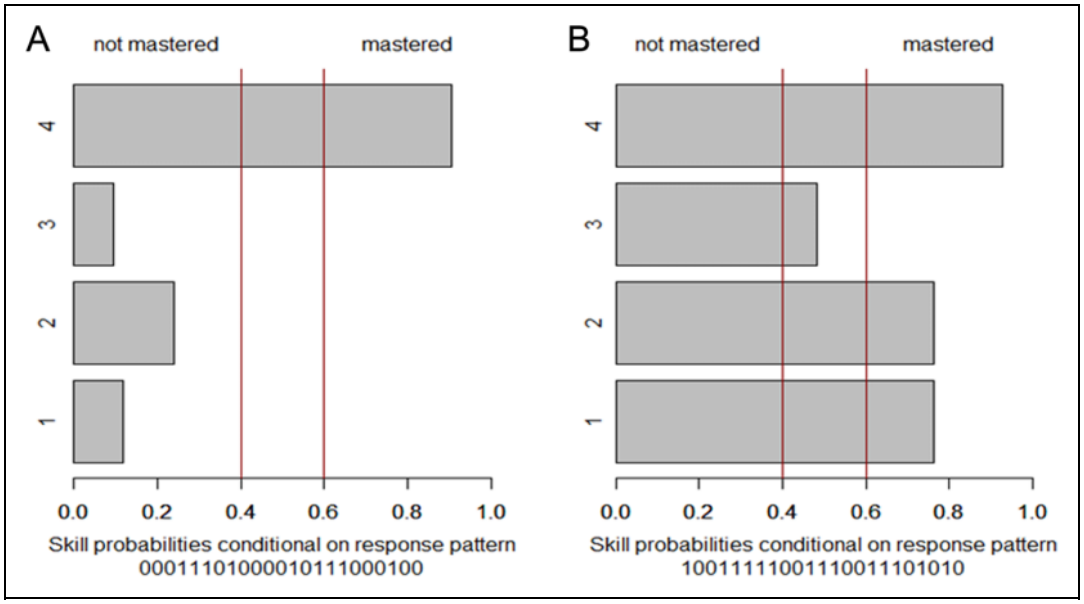

Figure 4. This figure depicts the probability of mastering each one of the attributes for two response patterns (A and B), resulting in a score of 9 and 14 in the 23-item test. The labels 1, . . . , 4 refer to each one of the attributes, namely, 1 = study habits, 2 = study attitudes, 3 = helping others, and 4 = generalized compliance.

Table 9. Relationship Among the SJT Sum Scores, the Logit Transformation of G-DINA Person Parameter Estimates, and the Criterion/Correlates.

									Attributes			
	GPA	RAVEN	NEU	EXT	OPE	AGR	CON	SJT	1	2	3	4
SJT sum score	.30**	.02	−.11	.20*	.28**	.25**	.46**	1				
Attributes												
1. Study habits	.35**	.02	−.10	.28**	.16	.27**	.53**	.77**	1			
2. Study attitudes	.23*	.06	−.06	.07	.17*	.24*	.35**	.70**	.63**	1		
3. Helping others	.28**	.15	−.10	.24**	.24**	.24**	.49**	.71**	.82**	.63**	1	
4. Generalized compliance	.17	-.12	−.02	.00	−.01	.14	.09	.38**	.29**	.38**	.02	1

Note: $N = 137$ when GPA is included in the comparison. GPA = grade point average; NEU = neuroticism; EXT = extraversion; OPE = openness; AGRE = agreeableness; CON = conscientiousness; SJT = situational judgment test; G-DINA = *generalized deterministic inputs, noisy "and" gate.*
*$p < .05$. **$p < .01$.

these personality measures. Generalized compliance (attribute 4) was not significantly related to any of the theoretically relevant variables. Although most attributes were highly intercorrelated, this was also not the case for generalized compliance (attribute 4). This attribute was not significantly related to helping others (attribute 3), and the correlations with the other attributes were moderate in size. Finally, note that neither the SJT sum score nor the attributes were significantly related to the RAVEN score (which might be due to the range restricted nature of the university student sample; see Peeters & Lievens, 2005).

Reliability of the assessment. The alpha reliability coefficient depends on the assumption that all the items reflect a single construct (Miller, 1995). Given that SJT items are typically heterogeneous, coefficient alpha can be expected to be an inaccurate measure of the true reliability (see Catano

et al., 2012). Indeed, the internal consistency of the SJT scores (.57) in this sample was rather low. As noted previously, it therefore makes sense to use a reliability coefficient that takes into consideration the multidimensional nature of the SJT items. More importantly, from the CTT, we cannot estimate the reliability for the underlying dimensions that are being measured by the SJT. CDMs represent a new approach for assessing the reliability of these scores. A common indicator of reliability in CDM is called attribute classification accuracy, which indicates how accurately a CDM classifies test takers into correct attribute profiles.

To estimate attribute classification accuracy, we use the calibrated model to generate simulated data so that we could study the attribute classification accuracy once the true classifications are known. For this purpose, the responses from 1,600 examinees were simulated, that is 100 examinees for each one of the $2^4 = 16$ possible attribute profiles (i.e., latent classes). The model employed was the G-DINA model, and the values of the item parameters were those estimated from the empirical data. Then we fitted the G-DINA model to the simulated data set. The following cut-off points were applied to the EAP estimates: We define *mastery* as a posterior probability of mastering the attribute above .50 and *nonmastery* as a probability between below .50. We calculated the proportion of times that a test taker is classified correctly according to the true classifications. This analysis allowed determining: (a) the *attribute level classification accuracy*, that is, the proportion of correct classifications for each of the four attributes, and (b) the *pattern level classification accuracy*, which is defined as the proportion of times that a test taker is correctly classified in all the assigned attributes.

Results of this simulation study show that the attribute level classification accuracy was considerably high. The proportion of correctly classified individual attributes was always at least .85 (.95, .93, .85, and .93 for attributes 1, 2, 3, and 4, respectively). With regard to the pattern level classification accuracy, the proportion of times all the classified attributes were classified correctly was also considerably high (76%). Regarding the proportion of times that a test taker was correctly classified at least in 2 or 3 attributes, the proportions increased to .94 and .97, respectively.

Discussion

Contributions of Cognitive Diagnosis Models

To date, in the SJT domain, some of the sources of validity (those based on internal structure and response processes) and reliability have not been appropriately addressed. Therefore, it has been reiterated that the constructs SJTs measure are unknown (e.g., Christian et al., 2010; Ployhart & Weekley, 2006). This article posited that the absence of an appropriate psychometric approach has been a major obstacle to move the field forward because traditional psychometric procedures (e.g., factor analysis and Cronbach's alpha) cannot deal with the item multidimensionality in SJTs.

In this study, we explored how the CDM approach can offer useful solutions to these predicaments. We illustrated how common validity and reliability concerns in SJT research can be addressed by assessing the Q-matrix specification, the model fit, and the item and examinee parameter estimates. As summarized in the following, we demonstrated that the advantages of CDM over CTT in providing a greater wealth of information in analyzing SJTs are fourfold.

First, we showed that the application of a CDM model allows getting a better understanding of the underlying internal structure of the SJT. In our empirical example, successful completion of the SJT was found to require four attributes: study habits, study attitudes, helping others, and generalized compliance. As we have seen, all of these attributes are positively correlated, except helping others and generalized compliance. Importantly, the empirical validation of the Q-matrix allows for the experts' decisions and judgments to be verified. This empirical validation of the Q-matrix resulted in a new specification for one item that was supported by substantive theory as well as increased the

item's discrimination power. On the basis of increased insight in the underlying multidimensional structure of the SJT, CDMs allow for separately scoring the different attributes that are measured by the test, which is not possible with the typical use of a single overall score in SJTs.

Second, CDMs can illuminate response processes underlying SJTs because they show which set of KSAOs are required for solving SJT items and whether or not one KSAO can potentially compensate for the others. Through the study of the model fit, we were able to determine that the G-DINA model achieved the best fit to the data, and constraining the model to be conjunctive or disjunctive (i.e., using the DINA and DINO models) led to a significant loss of fit. According to the item parameters, different types of processes were involved within the same test.[3] In the case of some items (e.g., Item 23), only test takers who have mastered all the required attributes had a high probability of selecting the most effective answer. In the case of other items (e.g., Item 8), the mastery of one or more attributes could make up for lack of mastery in other attributes. There were still other items (e.g., Item 5) in which mastering each of the attributes led to an increase in the probability of success on a certain item, whereas the effect of the interaction among the attributes was negligible.

Third, we showed how CDM can provide information about the relationships of the four underlying dimensions (attributes in CDM language) in the SJT and theoretically relevant variables. As expected, student-related attributes (study habits and attitudes) were significantly related to GPA (Aquino, 2011) and conscientiousness (Barrick & Mount, 1991; Hough et al., 1990; Hurtz & Donovan, 2000; Salgado, 1997; Tett et al., 1991), and the helping others attribute was significantly related to personality (Hough, 1992; LePine & Van Dyne, 2001; Mount et al., 1998; Organ & Ryan, 1995). In this way, when we model the multidimensional nature of SJT, we gain insights into the relationships among the SJT scores and theoretically relevant variables. This also signals which attributes do not function as expected, which might trigger efforts to redesign the test at hand. Contrary to prior research (Konovsky & Organ, 1996), for instance, generalized compliance was not significantly related to any of the variables. We tentatively attribute this result to a poor representation of the construct domain of generalized compliance. There were only six items measuring this attribute, and inspection of their item content revealed that all of them represented situations in which students had to follow the norms proposed by their teacher (e.g., stick with the existing timetable). Other aspects of the generalized compliance construct such as punctuality and not wasting time were not represented in the current items.

Fourth, we illustrated how CDMS can allow for the reliability of SJT scores to be studied from an angle different from how it is traditionally done (i.e., based on Cronbach's alpha or test-retest procedures). Test precision in CDM is similar to the logic underlying CTT. In many testing contexts, it is necessary to classify respondents into performance categories. Decision accuracy refers to the extent to which classifications based on the observed scores agree with the classifications based on the true scores. Similarly, classification accuracy in CDM is intended to measure the degree to which classifications based on observed scores matched the true attribute profile. In our empirical example, the agreement-rate calculation between true and estimated attribute profiles based on the simulated data indicated that the proportion of times that the entire attribute profile is recovered was considerably high. In addition, CDM results provided information about individual attribute classification accuracy. This enables researchers to determine whether any of the attributes was measured with low reliability. Taking the items with a high discrimination index as an example, additional assessment tasks could be designed, specifically for attributes with lower accuracy classification rates, so that the resulting SJT might achieve higher levels of reliability. These new items can be added to the calibrated item pool through linking designs, as it is often done in IRT. In the most common scenario, a group of examinees will take a set of old (i.e., calibrated) items and a set of new (i.e., uncalibrated) items.

Finally, apart from the fourfold information that test users and designers could get, CDMs also provide finer-grained information about test takers' strengths and weaknesses. This information could be fruitfully used by HR practitioners in SJT applications, such as personnel selection and needs analyses in training programs (Weekley et al., 2015). A generic example of the prototypical feedback was shown in the empirical example. That is, the feedback consists of a list of attributes and indicates per attribute the probability that the test taker has mastered the attribute. Providing this feedback to test takers is relatively straightforward. The main point to consider when making a decision on which cut-off point to employ to convert these probabilities into profiles is the goal of the assessment (e.g., the willingness to report low-reliable profiles). If all respondents must be classified one way or another, one can employ .50 as cut-off score. On the other hand, in some applied contexts, one might be more interested in selecting high-performing (e.g., personnel selection) or low-performing (e.g., educational assessment) individuals. If that is the case, one needs to ensure that those specific patterns are accurately estimated. In addition, cognitive diagnosis computer adaptive assessments (CD-CAT) serve as one possible solution for the problem of having nonclassified individuals (for an overview, see e.g., Huebner, 2010). The termination criterion is generally based on the accuracy with which the respondents are assessed. Thus, for example, the diagnostic assessment can only be terminated when the posterior probability that a respondent belongs to a given state (i.e., mastery or nonmastery) achieves an acceptable value (e.g., less than .20 or greater than .80).

Caveats Related to Cognitive Diagnosis Models

Some caveats related to CDM should be acknowledged. First, we want to emphasize that the initial list of attributes should be carefully developed. As noted, this can be done via a variety of methods such as prior research, theory, job analytic information, and think-aloud protocols. It is equally pivotal to verify the Q-matrix developed (de la Torr, & Chiu, 2015), as we did in our empirical example, to correct possible misspecifications in the original Q-matrix. De la Torre and Chiu (2015) showed that the empirical validation procedure can accurately identify and correct misspecified q-entries without altering correct entries, particularly when high-quality items are involved. This is typically the case in educational assessment where items tend to be highly discriminating, but the results cannot be directly extrapolated in the case of poor-quality items. Thus, we stress the importance of relying on the expert ratings to examine these discrepancies. We also suggest doing a cross-validation in another sample to avoid the possibility of capitalization on chance, which might bias the statistical estimates.

Second, the relations between CDM and CTT deserve attention. There are various points in common between these two approaches. Lee, de la Torre, and Park (2011) explored the relationships between CDM, CTT, and IRT indices. The pattern of correlations among CTT and CDM indices that they reported is very similar to the one we obtained: Difficulty and discrimination CTT and CDM indices are typically highly correlated. We do not see this similarity in results as a limitation of CDM. Rather, it is a positive point that specific CDM indices correspond to the results of CTT indices. Our results indicate that items can provide diagnostic information (e.g., help differentiate between respondents who have mastered more attributes and respondents who have mastered fewer attributes) even if they are not developed under a CDM framework. The CTT discrimination indices may provide guidance on the diagnostic value of an item. In this way, items with low corrected point-biserial correlation can be expected to have low discrimination in CDM. In addition, as shown in our article, CDM indices provide a host of extra information over and above CTT indices. One difference between CDMs and CTT, which is a potential disadvantage of CDMs, is that their parameters must be estimated. Standard error of model parameters can be used as a measure of the precision of the estimate. Standard error estimates depend on the sample size: As sample size

increases, the standard error decreases. Note, however, that it has been shown that when the model fits the data, the DINA model parameters are invariant (de la Torre & Lee, 2010). Thus, no matter what sample of respondents takes the test, the item parameter estimates will generally be the same. This means that item parameter estimates have to be estimated only once, provided the sample is representative of the population.

A third caveat related to the application presented in the current study is that the specification of Q-matrix was done after the test was developed. This approach, referred to as retrofitting, is actually commonly found in the CDM literature. A good example is the study of Templin and Henson (2006), who demonstrated how the hypothesized underlying factors contributing to pathological gambling can be measured with the DINO model. However, in those applications, where CDM have been retrofitted to assessments constructed using a unidimensional or CTT framework, convergence problems may occur, as well as poor item, respondent, or model fit (Rupp & Templin, 2008b). Thus, a more optimal approach is to design a test from the beginning and apply these theory-based specifications during the test development process itself (de la Torre, Tjoe, Rhoads, & Lam, 2010).

Conclusion

This study proposed and illustrated how CDM can be used to explore the nature of the constructs that SJTs measure, which is one of the current and principal challenges in SJT research (Ryan & Ployhart, 2014; Weekley et al., 2015). Overall, we conclude that CDMs include a greater wealth of information in analyzing SJTs than traditional procedures based on CTT do. That is, CDM holds promise in evaluating the internal structure of the SJT, providing information about the cognitive processes underlying the responses in the SJT, clarifying how and why the SJT scores relate to other variables, and leading to a more appropriate estimation of the reliability of these scores.

Acknowledgements

The authors wish to thank associate editor Adam Meade and three anonymous reviewers for their valuable comments and suggestions on earlier versions of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was partially supported by Grant PSI2013-44300-P (Ministerio de Economia y Competitividad and European Social Fund).

Notes

1. Based on the probabilities of being classified into an attribute profile given the data (i.e., $P(\boldsymbol{\alpha}_l|\boldsymbol{X})$), the individual attribute profile can be deduced via three methods: maximum likelihood estimation (MLE), maximum a posteriori (MAP) estimation, and expected a posteriori (EAP) estimation. For a comparison among MLE, MAP, and EAP classification methods, see Huebner and Wang (2011).
2. Currently there are different programs available for estimating cognitive diagnosis models (CDMs), for example the G-DINA framework in Ox (Doornik, 2002) by de la Torre, the MDLTM program by von Davier (2005), the LCDM framework in SAS (SAS Institute Inc., 2007) and Mplus (Muthén & Muthén, 2012) by Templin, Henson, Douglas, and Homan. The main advantage of R is that it is freely available and very flexible.

3. When referring to a particular underlying latent structure and the response processes implied, it should be acknowledged that between-subjects conclusions should not be interpreted at the individual level (Borsboom, Mellenbergh, & van Heerden, 2003). Recently, this issue has been considered in measurement equivalence (Tay, Meade, & Cao, 2015).

References

Adams, R. J., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23. doi:10.1177/0146621697211001

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automated Control*, *19*, 716-723. doi:10.1109/TAC.1974.1100705

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Aquino, L. B. (2011). Study habits and attitudes of freshmen students: Implications for academic intervention programs. *Journal of Language Teaching & Research*, *2*(5), 1116-1121. doi:10.4304/jltr.2.5.1116-1121

Arthur, W., Jr., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology*, *99*(3), 535-545. doi:10.1037/a0035788

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*(1), 1-26. doi:10.1111/j.1744-6570.1991.tb00688.x

Bateman, T. S., & Organ, D. W. (1983). Job satisfaction and the good soldier: The relationship between affect and employee "citizenship." *The Academy of Management Journal*, *26*(4), 587-595. doi:10.2307/255908

Bess, T. L., & Mullins, M. E. (2002, April). *Exploring a dimensionality of situational judgment: Task and contextual knowledge*. Paper presented at the 17th annual conference of the Society for Industrial and Organizational Psychology, Toronto, Canada.

Borman, W. C., White, L. A., & Dorsey, D. W. (1995). Effects of ratee task performance and interpersonal factors on supervisor and peer performance ratings. *Journal of Applied Psychology*, *80*(1), 168-177. doi:10.1037/0021-9010.80.1.168

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*, 203-219. doi:10.1037/0033-295X.110.2.203

Buck, G., VanEssen, T., Tatsuoka, K., Kostin, I., Lutz, D., & Phelps, M. (1998). *Development, selection and validation of a set of cognitive and linguistic attributes for the SAT I Verbal: Analogy section* (Research Report, RR-98-19). Princeton, NJ: Educational Testing Service.

Catano, V. M., Brochu, A., & Lamerson, Ch. D. (2012). Assessing the reliability of situational judgment tests used in high-stakes situations. *International Journal of Selection and Assessment*, *20*(3), 333-346. doi:10.1111/j.1468-2389.2012.00604.x

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*(2), 123-140. doi:10.1111/j.1745-3984.2012.00185.x

Christian, M., Edwards, B., & Bradley, J. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, *63*, 83-117. doi:10.1111/j.1744-6570.2009.01163.x

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Schmidt-Harvey, V. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology*, *86*, 410-417. doi:10.1037/0021-9010.86.3.410

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Credé, M., & Kuncel, N. R. (2008). Study habits, skills, and attitudes: The third pillar supporting collegiate academic performance. *Perspectives on Psychological Science*, *3*(6), 425-453. doi:10.1111/j.1745-6924.2008.00089.x

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*, 343-362. doi:10.1111/j.1745-3984.2008.00069.x

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179-199. doi:10.1007/s11336-011-9207-7

de la Torre, J., & Chiu, C. -Y. (2015). A general method of empirical Q-matrix validation. *Psychometrika*, 1-21. doi:10.1007/s11336-015-9467-8

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement*, *47*, 227-249. doi:10.1111/j.1745-3984.2010.00110.x

de la Torre, J., & Lee, Y. -S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement*, *47*(1), 115-127. doi:10.1111/j.1745-3984.2009.00102.x

de la Torre, J., Tjoe, H., Rhoads, K., & Lam, T. C. (2010, April). *Conceptual and theoretical issues in proportional reasoning*. Paper presented at Annual Meeting of American Educational Research Association, Denver, CO.

Doignon, J. P., & Falmagne, J. C. (1999). *Knowledge spaces*. Berlin: Springer.

Doornik, J. A. (2002). *Object-oriented matrix programming using Ox*. London: Timberlake Consultants Press.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359-374. doi:10.1016/0001-6918(73)90003-6

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological 15 Bulletin*, *76*(5), 378-382. doi:10.1037/h0031619

García, P. E., Olea, J., & de la Torre, J. (2014). Application of cognitive diagnosis models to competency-based situational judgment tests. *Psicothema*, *3*, 372-377. doi:10.7334/psicothema2013.322

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 333-352. doi:10.1111/j.1745-3984.1989.tb00336.x

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*, 262-277. doi:10.1177/0146621604272623

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65-70

Hough, L. M. (1992). The "Big Five" personality variables-construct confusion: Description versus prediction. *Human Performance*, *5*(1-2), 139-155. doi:10.1080/08959285.1992.9667929

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology*, *75*, 581-595. doi:10.1037/0021-9010.75.5.581

Huebner, A. (2010). An overview in recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research & Evaluation*, *15*, 1-7.

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, *71*, 407-419. doi:10.1177/0013164410388832

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology*, *85*(6), 869-879. doi:10.1037/0021-9010.85.6.869

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. Unpublished doctoral dissertation, University of Illinois, Champaign, IL.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258-272. doi:10.1177/01466210122032064

Kamdar, D., & Van Dyne, L. (2007). The joint effects of personality and workplace social exchange relationships in predicting task performance and citizenship performance. *Journal of Applied Psychology*, *92*(5), 1286-1298. doi:10.1037/0021-9010.92.5.1286

Konovsky, M. A., & Organ, D. W. (1996). Dispositional and contextual determinants of organizational citizenship behavior. *Journal of Organizational Behavior*, *17*(3), 253-266.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174. doi:10.2307/2529310

Lee, Y.-S., de la Torre, J., & Park, Y. S. (2011). Relationships between cognitive diagnosis, CTT, and IRT indices: An empirical investigation. *Asia Pacific Education Review*, *13*(2), 333-345. doi:10.1007/s12564-011-9196-3

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, UK: Cambridge University Press.

LePine, J. A., & Van Dyne, L. (2001). Voice and cooperative behavior as contrasting forms of contextual performance: Evidence of differential relationships with Big Five personality characteristics and cognitive ability. *Journal of Applied Psychology*, *86*(2), 326-336. doi:10.1037/0021-9010.86.2.326

Li, H., & Suen, H. K. (2013). Constructing and validating a Q-Matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment*, *18*(1), 1-25. doi:10.1080/10627197.2013.761522

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgement tests: A review of recent research. *Personnel Review*, *37*(4), 426-441. doi:10.1108/00483480810877598

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, *86*, 730-740. doi:10.1037/0021-9010.86.4.730

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, *60*(1), 63-91. doi:10.1111/j.1744-6570.2007.00065.x

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling*, *2*(3), 255-273. doi:10.1080/10705519509540013

Motowildo, S. J., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human Performance*, *10*(2), 71-83. doi:10.1207/s15327043hup1002_1

Mount, M. K., Barrick, M. R., & Stewart, G. L. (1998). Five-Factor Model of personality and performance in jobs involving interpersonal interactions. *Human Performance*, *11*(2-3), 145-165. doi:10.1080/08959285.1998.9668029

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.

Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome* (*Vol. xiii*). Lexington, MA: Lexington Books/D. C. Heath and Com.

Organ, D. W., & Ryan, K. (1995). A meta-analytic review of attitudinal and dispositional predictors of organizational citizenship behavior. *Personnel Psychology*, *48*(4), 775-802. doi:10.1111/j.1744-6570.1995.tb01781.x

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology*, *89*, 187-207. doi:10.1037/0021-9010.89.2.187

Patterson, F., Ashworth, V., Zibarras, L., Coan, P., Kerrin, M., & O'Neill, P. (2012). Evaluations of situational judgment tests to assess non-academic attributes in selection. *Medical Education*, *46*, 850-868. doi:10.1111/j.1365-2923.2012.04336.x

Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement*, *65*(1), 70-89. doi:10.1177/0013164404268672

Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment*, *11*(1), 1-16. doi:10.1111/1468-2389.00222

Ployhart, R. E., & Weekley, J. A. (2006). Situational judgment: Some suggestions for future science and practice. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 345-350). Mahwah, NJ: Erlbaum.

Proctor, B. E., Prevatt, F. F., Adams, K. S., Reaser, A., & Petscher, Y. (2006). Study skills profiles of normal-achieving and academically-struggling college students. *Journal of College Student Development*, *47*(1), 37-51. doi:10.1353/csd.2006.0011

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Advanced progressive matrices manual*. Oxford, UK: Oxford Psychologists Press.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2015). *CDM: Cognitive diagnosis modeling* (R package version 4.4 1). Retrieved from http://CRAN.R-project.org/package=CDM

Rupp, A. A. (2007). The answer is in the question: A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models. *International Journal of Testing*, *7*, 95-125. doi:10.1080/15305050701193454

Rupp, A. A., & Mislevy, R. J. (2007). Cognitive foundations of structured item response theory models. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment in education: Theory and practice* (pp. 205-241). Cambridge, UK: Cambridge University Press.

Rupp, A. A., & Templin, J. (2008a). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*(1), 78-96. doi: 10.1177/0013164407301545

Rupp, A. A., & Templin, J. L. (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, *6*(4), 219-262. doi:10.1080/15366360802490866

Ryan, A. M., & Ployart, E. (2014). A century of selection. *Annual Review of Psychology*, *65*, 693-717. doi:10.1146/annurev-psych-010213-115134

Salgado, J. F. (1997). The Five Factor Model of personality and job performance in the European Community. *Journal of Applied Psychology*, *82*(1), 30-43. doi:10.1037/0021-9010.82.1.30

SAS Institute Inc. (2007). *User's guide for SAS software navigator*. Cary, NC: Author.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*(2), 199-223. doi:10.1037/1082-989X.1.2.199

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests. Theory, measurement, and application* (pp. 135-155). Mahwah, NJ: Erlbaum.

Smith, C. A., Organ, D. W., & Near, J. P. (1983). Organizational citizenship behavior: Its nature and antecedents. *Journal of Applied Psychology*, *68*, 655-663. doi:10.1037/0021-9010.68.4.653

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconception based on item response theory. *Journal of Education Statistic*, *20*, 345-354. doi:10.1111/j.1745-3984.1983.tb00212.x

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*, 3-46. doi:10.1177/1094428114553062

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287-305. doi:10.1037/1082-989X.11.3.287

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, *44*(4), 703-742. doi:10.1111/j.1744-6570.1991.tb00696.x

von Davier, M. (2005). mdltm—multidimensional discrete latent trait modeling software [Computer software]. Princeton, NJ: Educational Testing Service

Weekley, J. A., Hawkes, B., Guenole, N., & Ployhart, R.E. (2015). Low-fidelity simulations. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*, 295-322. doi:10.1146/annurev-orgpsych-032414-111304

Weekley, J. A., & Ployhart, R. E. (2005). Situational judgement: Antecedents and relationships with performance. *Human Performance*, *18*, 81-104. doi:10.1207/s15327043hup1801_4

Weekley, J. A., & Ployhart, R. E. (2006). An introduction to situational judgment testing. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests* (pp. 1-10). Mahwah, NJ: Erlbaum.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review*, *19*, 188-202. doi:10.1016/j.hrmr.2009.03.007

Willse, J. T. (2014). *CTT: Classical test theory functions* (R package version 2.1). Retrieved from http://CRAN.R-project.org/package=CTT

Author Biographies

Miguel A. Sorrel is a PhD candidate at the Department of Social Psychology and Methodology at Universidad Autónoma de Madrid. His research interests include item response theory, cognitive diagnosis modeling, and computerized adaptive testing.

Julio Olea is a professor of methodology of behavioural sciences at Universidad Autónoma de Madrid. His teaching and research work has focused on the field of psychometric: implementation of scaling methods, cognitive diagnosis modeling, and computerized adaptive testing.

Francisco J. Abad is an associate professor at Universidad Autónoma de Madrid. He has developed his teaching and research work in the field of psychometrics: polytomous IRT models, goodness of fit, software development, and computerized adaptive testing.

Jimmy de la Torre is a professor of educational psychology at Rutgers University. His primary research interests are in the field of psychological and educational testing and measurement, particularly in the areas of item response theory, cognitive diagnosis modeling, and psychometric models for noncognitive test data.

David Aguado is an associate professor at the Universidad Autónoma de Madrid. He is the director of innovation in talent management at the Instituto de Ingeniería del Conocimiento (IIC). He has developed his teaching and research work in competency development and performance management.

Filip Lievens is full professor at the Department of Personnel Management, Work and Organizational Psychology at Ghent University. His research has influenced a variety of applied measurement issues in I-O psychology. He has published in the *Annual Review of Psychology*, *Journal of Applied Psychology*, *Personnel Psychology*, and *Journal of Management*.

Chapter 3

Inferential Item-Fit Evaluation in Cognitive Diagnosis Modeling

This study was published in the journal *Applied Psychological Measurement* (Sorrel, Abad, Olea, de la Torre, & Barrada, 2017). The following pages include this publication. This article was first published on May 19, 2017. The issue was published on November 1, 2017. The article can be downloaded from https://doi.org/10.1177/0146621617707510. The journal impact factor and 5-year impact factor (2016) are 0.885 and 1.291, respectively.

Article

Inferential Item-Fit Evaluation in Cognitive Diagnosis Modeling

Applied Psychological Measurement
2017, Vol. 41(8) 614–631

DOI: 10.1177/0146621617707510
journals.sagepub.com/home/apm
SAGE

Miguel A. Sorrel[1], Francisco J. Abad[1], Julio Olea[1], Jimmy de la Torre[2], and Juan Ramón Barrada[3]

Abstract

Research related to the fit evaluation at the item level involving cognitive diagnosis models (CDMs) has been scarce. According to the parsimony principle, balancing goodness of fit against model complexity is necessary. General CDMs require a larger sample size to be estimated reliably, and can lead to worse attribute classification accuracy than the appropriate reduced models when the sample size is small and the item quality is poor, which is typically the case in many empirical applications. The main purpose of this study was to systematically examine the statistical properties of four inferential item-fit statistics: $S - X^2$, the likelihood ratio (LR) test, the Wald (W) test, and the Lagrange multiplier (LM) test. To evaluate the performance of the statistics, a comprehensive set of factors, namely, sample size, correlational structure, test length, item quality, and generating model, is systematically manipulated using Monte Carlo methods. Results show that the $S - X^2$ statistic has unacceptable power. Type I error and power comparisons favor LR and W tests over the LM test. However, all the statistics are highly affected by the item quality. With a few exceptions, their performance is only acceptable when the item quality is high. In some cases, this effect can be ameliorated by an increase in sample size and test length. This implies that using the above statistics to assess item fit in practical settings when the item quality is low remains a challenge.

Keywords

cognitive diagnosis models, item-fit statistics, absolute fit, relative fit, Type I error, power

Cognitive diagnosis models (CDMs) have been actively researched in the recent measurement literature. CDMs are multidimensional, and confirmatory models specifically developed to identify the presence or absence of multiple attributes involved in the assessment items (for an overview of these models, see, for example, DiBello, Roussos, & Stout, 2007; Rupp & Templin, 2008). Although originally developed in the field of education, these models have been used in measuring

[1]Universidad Autónoma de Madrid, Spain
[2]The University of Hong Kong, Hong Kong
[3]Universidad de Zaragoza, Spain

Corresponding Author:
Miguel A. Sorrel, Department of Social Psychology and Methodology, Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, Madrid 28049, Spain.
Email: miguel.sorrel@uam.es

other types of constructs, such as psychological disorders (e.g., de la Torre, van der Ark, & Rossi, 2015; Templin & Henson, 2006) and situation-based competencies (Sorrel et al., 2016).

There are currently no studies comparing item characteristics (e.g., discrimination, difficulty) as a function of the kind of the constructs being assessed. However, some data suggest that important differences can be found. Specifically, notable differences are found for item discrimination, which is one of the most common indices used to assess item quality. Item discrimination relates to how well an item can accurately distinguish between respondents who differ on the constructs being measured. Although it does not account for the attribute complexity of the items, a simple measure of discrimination is defined as the difference between the probabilities of correct response for those respondents mastering all and none of the required attributes. This index is bounded by 0 and 1. In empirical applications, such as the fraction subtraction data described and used by Tatsuoka (1990) and by de la Torre (2011), one of the most widely used datasets in CDM in the educational context, the mean discrimination power of the items was 0.80. In contrast, when CDMs have been applied in applications outside educational measurement, the resulting discrimination estimates were found to be in the 0.40 range (de la Torre et al., 2015; H.-Y. Liu, You, Wang, Ding, & Chang, 2013; Sorrel et al., 2016; Templin & Henson, 2006). In these empirical applications, researchers typically used a sample size that varies approximately from 500 (e.g., de la Torre, 2011; Templin & Henson, 2006) to 1,000 (de la Torre et al., 2015), with an average number of items equal to 30, 12 being the minimum (de la Torre, 2011). Different CDMs were considered, including the deterministic inputs, noisy ''and'' gate (DINA; Haertel, 1989) model; the deterministic inputs, noisy ''or'' gate (DINO) model (Templin & Henson, 2006); the additive CDM (*A*-CDM; de la Torre, 2011); and the generalized deterministic inputs, noisy ''and'' gate (G-DINA; de la Torre, 2011) model.

Given the large number of different models, one of the critical concerns in CDM is selecting the most appropriate model from the available CDMs. Each CDM assumes a specified form of item response function (IRF). In the CDM context, the IRF denotes the probability that an item j is answered correctly as a function of the latent class. This study focused on methods assessing this assumption. Model fit evaluated at the test level simultaneously takes all the items into consideration. However, when there is model–data misfit at the test level, the misfit may be due to a (possibly small) subset of the items. Item-level model fit assessment allows us to identify these misfitting items. The research focused on item fit is important because such analysis can provide guidelines to practitioners on how to refine a measurement instrument. This is a very important topic because current empirical applications reveal that no one single model can be used for all the test items (see, for example, de la Torre & Lee, 2013; de la Torre et al., 2015; Ravand, 2016). Consequently, in this scenario, item-fit statistics are a useful tool for selecting the most appropriate model for each item. The main purpose of this study was to systematically examine the Type I error and power of four-item-fit statistics, and provide information about the usefulness of these indexes across different plausible scenarios. Only goodness-of-fit measures with a significance test associated with them (i.e., inferential statistical evaluation) were considered in this article. The rest of the article is structured as follows: First is a brief introduction of the generalized DINA model framework. This is followed by a review of item-fit evaluation in CDM, and for a presentation of the simulation study designed to evaluate the performance of the different item-fit statistics. Finally, the results of the simulation study and the implications and future studies are discussed.

The Generalized DINA Model Framework

In many situations, the primary objective of CDM was to classify examinees into 2^K latent classes for an assessment diagnosing K attributes. Each latent class is represented by an

attribute vector denoted by $\boldsymbol{\alpha}_l = (\alpha_{l1}, \alpha_{l2}, \ldots, \alpha_{lK})$, where $l = 1,\ldots, 2^K$. All CDMs can be expressed as $P(X_j = 1/\boldsymbol{\alpha}_l) = P_j(\boldsymbol{\alpha}_l)$, the probability of success on item j conditional on the attribute vector l. For diagnostic purposes, the main CDM output of interest is the estimate of examinee i's $\boldsymbol{\alpha}_i = \{\alpha_{ik}\}$.

Several general models that encompass reduced (i.e., specific) CDMs have been proposed, which include the above-mentioned G-DINA model, the general diagnostic model (GDM; von Davier, 2005), and the log-linear CDM (LCDM; Henson, Templin, & Willse, 2009). In this article, the G-DINA model, which is a generalization of the DINA model, is used. The G-DINA model describes the probability of success on item j in terms of the sum of the effects of the attributes involved and their corresponding interactions. This model partitions the latent classes into $2^{K_j^*}$ latent groups, where K_j^* is the number of required attributes for item j. Each latent group represents one reduced attribute vector, $\boldsymbol{\alpha}_{lj}^*$, that has its own associated probability of success, written as

$$P(\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}, \quad (1)$$

where δ_{j0} is the intercept for item j, δ_{jk} is the main effect due to α_k, $\delta_{jkk'}$ is the interaction effect due to α_k, and $\alpha_{k'}$ and $\delta_{j12\ldots K_j^*}$ are the interaction effects due to $\alpha_1, \ldots, \alpha_{K_j^*}$. Thus, without constraints on the parameter values, there are $2^{K_j^*}$ parameters to be estimated for item j.

The G-DINA model is a saturated model that subsumes several widely used reduced CDMs, including the DINA model, the DINO model, the *A*-CDM, the linear logistic model (LLM; Maris, 1999), and the reduced reparametrized unified model (R-RUM; Hartz, 2002). Although based on different link functions, *A*-CDM, LLM, and R-RUM are all additive models, where the incremental probability of success associated with one attribute is not affected by those of other attributes. Ma, Iaconangelo, and de la Torre (2016) found that, in some cases, one additive model can closely recreate the IRF of other additive models. Thus, in this work, only three of these reduced models corresponding to the three types of condensation rules are considered: DINA model (i.e., conjunctive), DINO model (i.e., disjunctive), and the *A*-CDM (i.e., additive). If several attributes are required to correctly answer the items, the DINA model is deduced from the G-DINA model by setting all terms except for δ_{j0} and $\delta_{j12\ldots K_j^*}$ to 0. As such, the DINA model has two parameters per item. Likewise, the DINO model has two parameters per item, and can be obtained from the G-DINA model by setting $\delta_{jk} = -\delta_{jkk'} = \cdots = (-1)^{K_j^*+1}\delta_{j12\ldots K_j^*}$. When all the interaction terms are dropped, the G-DINA model under the identity link reduces to the *A*-CDM, which has $K_j^* + 1$ parameters per item. Each of these models assumes a different cognitive process in solving a problem (for a detailed description, see de la Torre, 2011).

Item-Fit Evaluation

The process of model selection involves checking the model–data fit, which can be examined at test, item, or person level. Extensive studies have been conducted to evaluate the performance of various fit statistics at the test level (e.g., Chen, de la Torre, & Zhang, 2013; Y. Liu, Tian, & Xin, 2016) and at the person level (e.g., Cui & Leighton, 2009; Y. Liu, Douglas, & Henson, 2009). At the item level, some item-fit statistics have also been recently proposed to evaluate absolute fit (i.e., the discrepancy between a statistical model and the data) and relative fit (i.e., the discrepancy between two statistical models). The parsimony principle dictates that from a group of models that fit equally well, the simplest model should be chosen. The lack of parsimony, or overfitting, may result in a poor generalization performance of the results to new

data because some residual variation of the calibration data is captured by the model. With this in mind, general CDMs should not be always the preferred model. In addition, as pointed out by de la Torre and Lee (2013), there are several reasons that make reduced models preferable to general models: First, general CDMs are more complex, thus requiring a larger sample size to be estimated reliably. Second, reduced models have parameters with a more straightforward interpretation. Third, appropriate reduced models lead to better attribute classification accuracy than the saturated model, particularly when the sample size is small and the item quality is poor (Rojas, de la Torre, & Olea, 2012). In this line, Ma et al. (2016) found that a combination of different appropriate reduced models determined by the Wald test always produced a more accurate classification accuracy than the unrestricted model (i.e., the G-DINA model). In the following, some of the statistics that may be computed in this context will be described.

Absolute Fit

Absolute item fit is typically assessed by comparing the item performance on various groups with the performance levels predicted by the fitted model. A χ^2-like statistic is used to make this comparison. Different statistics have emanated from traditional item response theory (IRT), and the main difference among them is how the groups are formed. There are two main approaches: In the first one, respondents are grouped based on their latent trait estimates, and observed frequencies of correct/incorrect responses for these groups are obtained. Yen's (1981) Q_1 statistic is computed using this approach and has been adapted to CDM (Sinharay & Almond, 2007; C. Wang, Shu, Shang, & Xu, 2015). Its performance has been compared with that of the posterior predictive model checking method (Levy, Mislevy, & Sinharay, 2009). Q_1 Type I error was generally well kept below .05 and was preferred to the posterior predictive model checking method. The main problem with this approach is that observed frequencies are not truly observed because they cannot be obtained without first fitting a certain model. This will lead to a model-dependent statistic that makes it difficult to determine the degrees of freedom (Orlando & Thissen, 2000; Stone & Zhang, 2003). In the second approach, the statistic is formulated based on the observed and expected frequencies of correct/incorrect responses for each summed score (Orlando & Thissen, 2000). The main advantage of this approach is that the observed frequencies are solely a function of observed data. Thus, the expected frequencies can be compared directly with observed frequencies in the data. A χ^2-like statistic, referred to as $S-X^2$ (Orlando & Thissen, 2000), is then computed as

$$S-X^2{}_j=\sum_{s=1}^{J-1}N_s\frac{\left(O_{js}-E_{js}\right)^2}{E_{js}\left(1-E_{js}\right)}\sim\chi^2(J-1-m), \tag{2}$$

where s is the score group, J is the number of items, N_s is the number of examinees in group s, and O_{js} and E_{js} are the observed and predicted proportions of correct responses for item j for group s, respectively. The model-predicted probability of correctly responding item j for examinees with sum score s is defined as

$$P(x_{ij}=1|S_i=s)=\frac{\sum_{l=1}^{2^K}P\left(x_{ij}=1|\boldsymbol{\alpha}_l\right)P\left(S_i^j=s-1|\boldsymbol{\alpha}_l\right)p(\boldsymbol{\alpha}_l)}{\sum_{l=1}^{2^K}P(S_i=s|\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)}, \tag{3}$$

where $P(S_i = s|\boldsymbol{\alpha}_l)$ is the probability of obtaining the sum score $s-1$ in the test composed of all the items except item j, and $p(\boldsymbol{\alpha}_l)$ defines the probability for each of the latent classes. Model-predicted joint-likelihood distributions for each sum score are computed using the recursive algorithm developed by Lord and Wingersky (1984), and detailed in Orlando and Thissen (2000). The statistic is assumed to be asymptotically χ^2 distributed with $J-1-m$ degrees of freedom, where m is the number of item parameters.

Relative Fit

When comparing different nested models, there are three common tests that can be used (Buse, 1982): likelihood ratio (LR), Wald (W), and Lagrange multiplier (LM) tests. In the CDM context, the null hypothesis (H_0) for these tests assumes that the reduced model (e.g., A-CDM) is the ''true'' model, whereas the alternative hypothesis (H_1) states that the general model (i.e., G-DINA) is the ''true'' model. As such, H_0 defines a restricted parameter space. For example, for an item j measuring two attributes in the A-CDM model, the interaction term is restricted to be equal to 0, whereas this parameter is freely estimated in the G-DINA model. It should be noted that the three procedures are asymptotically equivalent (Engle, 1983). In all the three cases, the statistic is assumed to be asymptotically χ^2 distributed with $2^{K_j^*}-p$ degrees of freedom, where p is the number of parameters of the reduced model.

Let $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ denote the maximum-likelihood estimates of the item parameters under H_0 and H_1, respectively (i.e., restricted and unrestricted estimates of the population parameter). Although all three tests answer the same basic question, their approaches to answering the question differ slightly. For instance, the LR test requires estimating the models under H_0 and H_1; in contrast, the W test requires estimating only the model under H_1, whereas the LM test requires estimating only the model under H_0.

Before describing in greater detail these three statistical tests, it is necessary to mention a few points about the estimation procedure in CDM. The parameters of the G-DINA model can be estimated using the marginalized maximum-likelihood estimation (MMLE) algorithm as described in de la Torre (2011). By taking the derivative of the log-marginalized likelihood of the response data, $l(\mathbf{X})$, with respect to the item parameters, $P_j(\boldsymbol{\alpha}_{lj}^*)$, the *estimating* function is obtained:

$$\frac{\partial l(\mathbf{X})}{\partial P_j\left(\boldsymbol{\alpha}_{lj}^*\right)} = \left[\frac{1}{P_j\left(\boldsymbol{\alpha}_{lj}^*\right)\left(1-P_j\left(\boldsymbol{\alpha}_{lj}^*\right)\right)}\right]\left[R_{\alpha_{lj}^*} - P_j\left(\boldsymbol{\alpha}_{lj}^*\right)I_{\boldsymbol{\alpha}_{lj}^*}\right], \tag{4}$$

where $I_{\boldsymbol{\alpha}_{lj}^*}$ is the number of respondents expected to be in the latent group $\boldsymbol{\alpha}_{lj}^*$, and $R_{\boldsymbol{\alpha}_{lj}^*}$ is the number of respondents in the latent group $\boldsymbol{\alpha}_{lj}^*$ expected to answer item j correctly. Thus, the MMLE estimate of $P_j(\boldsymbol{\alpha}_{lj}^*)$ is given by $\hat{P}_j(\boldsymbol{\alpha}_{lj}^*) = R_{\boldsymbol{\alpha}_{lj}^*} \big/ I_{\boldsymbol{\alpha}_{lj}^*}$. Estimating functions are also known as *score* functions in the LM context. The second derivative of the log-marginalized likelihood with respect to $P_j(\boldsymbol{\alpha}_{lj}^*)$ and $P_j(\boldsymbol{\alpha}_{l'j}^*)$ can be shown to be (de la Torre, 2011)

$$-\sum_{i=1}^{I}\left\{p\left(\boldsymbol{\alpha}_{lj}^*|\mathbf{X}_i\right)\frac{X_{ij}-P_j\left(\boldsymbol{\alpha}_{lj}^*\right)}{P_j\left(\boldsymbol{\alpha}_{lj}^*\right)\left[1-P_j\left(\boldsymbol{\alpha}_{lj}^*\right)\right]}\right\}\left\{p\left(\boldsymbol{\alpha}_{l'j}^*|\mathbf{X}_i\right)\frac{X_{ij}-P_j\left(\boldsymbol{\alpha}_{l'j}^*\right)}{P_j\left(\boldsymbol{\alpha}_{l'j}^*\right)\left[1-P_j\left(\boldsymbol{\alpha}_{l'j}^*\right)\right]}\right\}, \tag{5}$$

where $p(\boldsymbol{\alpha}_{lj}^*/\mathbf{X}_i)$ represents the posterior probability that examinee i is in latent group $\boldsymbol{\alpha}_{lj}^*$. Using $\hat{P}_j(\boldsymbol{\alpha}_{lj}^*)$ and the observed $\mathbf{X}$ to evaluate Equation 4, the information matrix is obtained for the

parameters of item j, $\mathbf{I}(\hat{\mathbf{P}}_j^*)$, and its inverse corresponds to the variance–covariance matrix, $Var(\hat{\mathbf{P}}_j^*)$, where $\hat{\mathbf{P}}_j^* = \{\hat{P}_j(\boldsymbol{\alpha}_{lj}^*)\}$ denotes the probability estimates.

LR test. As previously noted, the LR test requires the estimation of both unrestricted and restricted models. The likelihood function is defined as the probability of observing $\mathbf{X}$ given the hypothesis. It is defined as $L(\tilde{\boldsymbol{\theta}})$ for the null hypothesis and $L(\hat{\boldsymbol{\theta}})$ for the alternative hypothesis. The LR statistic is computed as twice the difference between the logs of the two likelihoods:

$$LR = 2\left[\log L\left(\hat{\boldsymbol{\theta}}\right) - \log L\left(\tilde{\boldsymbol{\theta}}\right)\right] \sim \chi^2\left(2^{K_j^*} - p\right), \tag{6}$$

where $\log L(\boldsymbol{\theta}) = \log \prod_{i=1}^{I} \sum_{l=1}^{L} L(\mathbf{X}_i|\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)$ and $L(\mathbf{X}_i|\boldsymbol{\alpha}_l) = \prod_{j=1}^{J} P(\boldsymbol{\alpha}_{lj})^{X_{ij}}[1 - P(\boldsymbol{\alpha}_{lj})^{1-X_{ij}}]$. Having a test composed of J items, the application of the LR test at the item level implies that $J_{K_j^*>1}$ comparisons will be made, where $J_{K_j^*>1}$ is the number of items measuring at least $K = 2$ attributes. For each of the $J_{K_j^*>1}$ comparisons, a reduced model is fitted to a target item, whereas the general model is fitted to the rest of the items. This model is said to be a restricted model because it has less parameters than an unrestricted model where the G-DINA is fitted to all the items. The LR test can be conducted to determine if the unrestricted model fits the data significantly better than the restricted model comparing the likelihoods of both the unrestricted and restricted models (i.e., $L(\hat{\boldsymbol{\theta}})$ and $L(\tilde{\boldsymbol{\theta}})$, respectively). Note that the likelihoods here are computed at the test level.

W test. The W test takes into account the curvature of the log-likelihood function, which is denoted by $C(\hat{\boldsymbol{\theta}})$, and defined by the absolute value of $\partial^2 \log L/\partial\boldsymbol{\theta}^2$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. In CDM research, de la Torre (2011) originally proposed the use of the W test to compare general and specific models at the item level under the G-DINA framework. For item j and a reduced model with p parameters, this test requires setting up $\mathbf{R}_j$, a $(2^{K_j^*} - p)\times 2^{K_j^*}$ restriction matrix with specific constraints that make the saturated model to be equivalent to the reduced model of interest. The W statistic is computed as

$$W_j = \left[\mathbf{R}_j \times \hat{\mathbf{P}}_j^*\right]' \left[\mathbf{R}_j \times Var\left(\hat{\mathbf{P}}_j^*\right) \times \mathbf{R}_j'\right]^{-1} \left[\mathbf{R}_j \times \hat{\mathbf{P}}_j^*\right] \sim \chi^2\left(2^{K_j^*} - p\right), \tag{7}$$

where $\hat{\mathbf{P}}_j^*$ are the unrestricted estimates of the item parameters.

LM test. The LM test is based on the slope of the log-marginalized likelihood $S(\boldsymbol{\theta}) = \partial \log L/\partial\boldsymbol{\theta}$, which is called the *score* function. By definition, $S(\boldsymbol{\theta})$ is equal to 0 when evaluated at the unrestricted estimates of $\boldsymbol{\theta}$ (i.e., $\hat{\boldsymbol{\theta}}$), but not necessarily when evaluated at the restricted estimates (i.e., $\tilde{\boldsymbol{\theta}}$). The score function is weighted by the information matrix to derive the LM statistics. Following the parameter estimation under the G-DINA framework, the score function can be assumed to be as indicated in Equation 4. The LM statistic for item j is defined as

$$LM_j = S_j(\tilde{\mathbf{P}}_j^*)' Var(\tilde{\mathbf{P}}_j^*) S_j(\tilde{\mathbf{P}}_j^*) \sim \chi^2\left(2^{K_j^*} - p\right), \tag{8}$$

where $\tilde{\mathbf{P}}_j^*$ are the restricted estimates of the item parameters. It should be noted that all item parameters are estimated under the restricted model.

Before these statistics can be used with real data, it must be ensured that they have good statistical properties. This is even more crucial for $S - X^2$, LR, and LM tests because they have not been examined before in the CDM context. There have been, however, noteworthy studies on $S - X^2$ in the IRT framework by Orlando and Thissen (2000, 2003) and Kang and Chen (2008). Its Type I error was generally found to be close to the nominal level. The LM test has also been applied within the IRT framework. It has been shown to be a useful tool for

evaluating the assumption of the form of the item characteristic curves in the two- and three-parameter logistic models (Glas, 1999; Glas & Suárez-Falcón, 2003). However, item quality was not manipulated in these previous studies, and its effect is yet to be determined. This factor has been found to be very relevant in many different contexts using the relative item-fit indices, as is the case of the evaluation of differential item functioning (DIF). For example, previous research using the LR test in DIF has found that the statistical power of the LR test to detect DIF increases with increases in item discrimination (W. C. Wang & Yeh, 2003).

The W test is the only one that has been used before in the CDM context for assessing fit at the item level. However, only two simulation studies examining their statistical properties were found. Although these works have contributed to our state of knowledge in this field, many questions related to the usefulness of these statistics with empirical data remained open. de la Torre and Lee (2013) studied the W test in terms of Type I error and power, and they found that it had a relative accurate Type I error and high power, particularly with large samples and items measuring a small number of attributes. In their case, the number of items was fixed to 30, and item quality was not manipulated. Items were set to have a mean discrimination power of approximately 0.60. Recently, Ma et al. (2016) extended the findings of de la Torre and Lee (2013) by including two additional reduced models (i.e., LLM and R-RUM). In their simulation design, they also considered two additional factors: item quality and attribute distribution. They found that, although item quality strongly influenced the Type I error and power, the effect of the attribute distribution (i.e., uniform or high order) was negligible. As a whole, although these studies have shed some light on the performance of the W test, the impact of other important factors or levels not explicitly considered in these studies remains unclear. This study aims to fill this gap, as well as examine the potential use of $S-X^2$, LR, and LM tests for item-fit evaluation in the CDM context.

Method

A simulation study was conducted to investigate the performance of several item-fit statistics. Five factors were varied, and their levels were chosen to represent realistic scenarios detailed in the introduction. These factors are as follows: (a) generating model (*MOD*; DINA model, *A*-CDM, and DINO model), (b) test length (*J*; 12, 24, and 36 items), (c) sample size (*N*; 500 and 1,000 examinees), (d) item quality or discrimination, defined as the difference between the maximum and the minimum probabilities of correct response according to the attribute latent profile (*IQ*; .40, .60, and .80), and (e) correlational structure (*DIM*; uni- and bidimensional scenarios).

The following are details of the simulation study: The probabilities of success for individuals who mastered none (all) of the required attributes were fixed to .30 (.70), .20 (.80), and .10 (.90) for the low, medium, and high item quality conditions, respectively. For the *A*-CDM, an increment of $.40 / K_j^*$, $.60 / K_j^*$, and $.80/ K_j^*$ was associated with each attribute mastery for the low, medium, and high item quality conditions, respectively. The number of attributes was fixed to $K = 4$. The correlational matrix of the attributes has an off-diagonal element of .5 in the unidimensional scenario, and 2×2 block diagonal submatrices with a correlation of .5 in the bidimensional scenario. The Q-matrices used in simulating the response data and fitting the models are given in Online Annex 1. There were the same number of one-, two-, and three-attribute items.

The $3 \times 3 \times 2 \times 3 \times 2$ ($MOD \times J \times N \times IQ \times DIM$) between-subjects design produces a total of 108 factor combinations. For each condition, 200 datasets were generated, and DINA, *A*-CDM, DINO, and G-DINA models were fitted. Type I error was computed as the proportion of times that H_0 was rejected when the fitted model is true. Power was computed as the proportion of times that a wrong reduced model is rejected. For example, in the case of the DINA model, power was computed as the proportion of times that H_0 was rejected when the generating model

is the *A*-CDM or the DINO model. Type I error and power were investigated using .05 as the significance level. With 200 replicates, the 95% confidence interval for Type I error is given by $.05 \pm 1.96\sqrt{.05(1-.05)/200} = [0.02, 0.08]$. For the purposes of this work, a power of at least .80 was considered adequate. The power analysis may not be interpretable when the Type I error for the statistics compared is very disparate. To make meaningful comparisons, it was necessary to approximate the distribution of the item-fit statistic under the null hypothesis. In doing so, the results from the simulation study were used. A nominal alpha (α_n) for which the actual alpha (α_a) was equal to .05 was found for all cases (i.e., simulation conditions of the design) where Type I error was either deflated or inflated (i.e., $\alpha_a \notin [.02, .08]$). In these cases, this adjusted value was used as α_n producing a value for power which could then be compared with the other statistical tests.

As a mean to summarize and better understand the results of the simulation study, separate ANOVAs were performed for each of the item-fit statistics. Dependent variables were Type I error and power associated with each statistical test for all items with the five factors as between-subjects factors. Due to the large sample size, most effects were significant. For this reason, omega square ($\hat{\omega}^2$), measure of effect size, was chosen to establish the impact of the independent variables. The following guidelines were considered for interpreting $\hat{\omega}^2$ (Kirk, 1996): Effect sizes in the intervals [0.010, 0.059), [0.059, 0.138), and [0.138, ∞) were considered small, medium, and large, respectively. In addition, a cutoff of $\hat{\omega}^2 \geq .138$ was used to establish the most salient interactions. It was checked that the estimates of observed power (i.e., post hoc power) were greater than .80. The code used in this article was written in R. Some functions included in the CDM (Robitzsch, Kiefer, George, & Uenlue, 2015) and G-DINA (Ma & de la Torre, 2016) packages were employed. The R code can be requested by contacting the corresponding author.

Results

Due to space constraints, only effect sizes are discussed and marginal means for the most relevant effects are reported. Type I error and power of the item-fit statistics for the three reduced models in their entirety are shown in Online Annexes 2 and 3.

Type I Error

The effect size $\hat{\omega}^2$ values and marginal means associated with each main effect on the Type I error are provided in Table 1. $S-X^2$ is the only statistic with a Type I error that was usually close to the nominal level. The marginal means are always within the [0.02, 0.08] interval, with the grand mean being 0.06. Only a small effect of item quality ($\hat{\omega}^2$ = .01) and the generating model ($\hat{\omega}^2$ = .03) was found: Type I error was slightly larger in the low and medium item quality conditions and for the *A*-CDM. None of the interactions had a salient effect.

The Type I errors of the LR, W, and LM tests were very similar. Type I error was only acceptable for the high item quality conditions, which was the factor with the greatest effect ($\hat{\omega}^2$ = .33, .71, and .30 for LR, W, and LM tests, respectively). When the item discrimination is low or medium, the Type I error was inflated. This makes it difficult to interpret the marginal means for all other factors, because conditions with low, medium, and high item discrimination are mixed. That was why the marginal means were generally much larger than the upper-limit of the confidence interval (i.e., 0.08). All things considered, the grand means of the three tests were inflated: 0.19, 0.29, and 0.14 for LR, W, and LM tests, respectively. Only one of the two-way interactions had a salient effect: Generating Model × Item Quality. As can be observed from Figure 1, there were large differences between the marginal means for the different levels of

Table 1. Marginal Means and Effect Sizes of the ANOVA Main Effects for the Type I Error.

	Data factor/level																	
	N			DIM			*J*				*IQ*				*MOD*				
Item-fit statistic	$\hat{\omega}^2$	500	1,000	$\hat{\omega}^2$	UNI	BI	$\hat{\omega}^2$	12	24	36	$\hat{\omega}^2$	LD	MD	HD	$\hat{\omega}^2$	DINA	*A*-CDM	DINO	Grand mean
$S - X^2$	.00	.07	.07	.00	.06	.07	.00	.06	.07	.07	**.01**	.07	.07	.06	**.03**	.06	.08	.06	0.06
LR	**.01**	.21	.18	.00	.19	.20	**.02**	.22	.19	.17	**.33**	.30	.23	.06	**.14**	.15	.17	.27	0.19
W	**.03**	.31	.27	.00	.29	.29	**.17**	.36	.28	.24	**.71**	.51	.29	.08	**.18**	.24	.27	.36	0.29
LM	**.01**	.15	.13	**.02**	.13	.15	**.03**	.16	.14	.13	**.30**	.16	.20	.07	**.60**	.16	.01	.26	0.14

Note. Effect size values greater than .010 are shown in bold. Shaded cells correspond to Type I error in the [0.02, 0.08] interval. *N* = sample size; DIM = correlational structure; *J* = test length; IQ = item quality; MOD = generating model; Uni = unidimensional; Bi = bidimensional; LD = low discrimination; MD = medium discrimination; HD = high discrimination; DINA = deterministic inputs, noisy and gate; CDM = cognitive diagnosis model; *A*-CDM = additive CDM; DINO = deterministic inputs, noisy "or" gate.

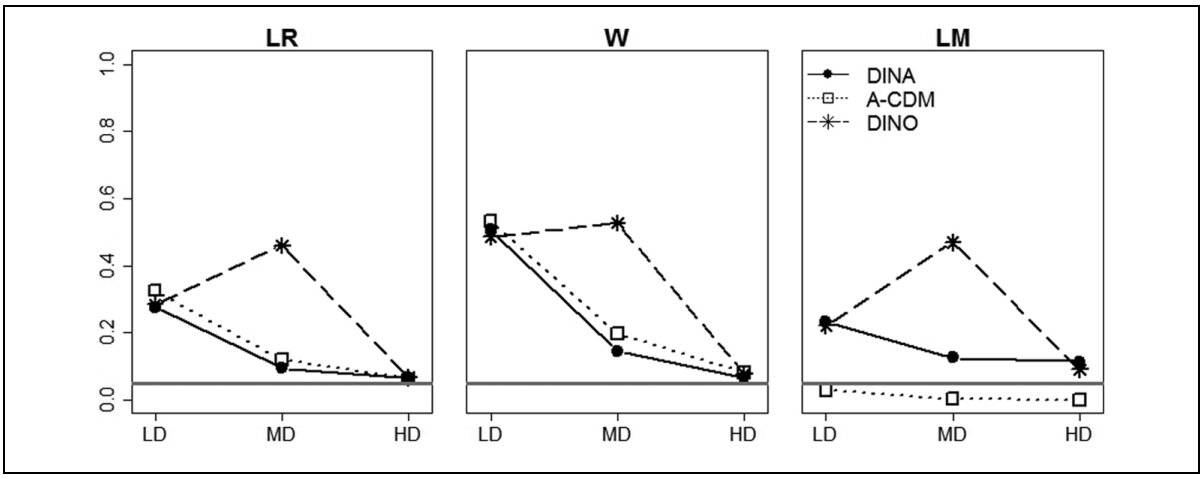

Figure 1. Two-way interaction of Generating Model × Item Quality with LR, W, and LM Type I error as dependent variables.

Note. The horizontal gray line denotes the nominal Type I error (α = .05). DINA = deterministic inputs, noisy and gate; CDM = cognitive diagnosis model; *A*-CDM = additive CDM; DINO = deterministic inputs, noisy "or" gate.

generating model across the levels of item quality. The Type I error was closer to the nominal level when item quality got higher, with the exception of the DINO model, where Type I error was more inflated with medium-quality items. Marginal means for the high-quality conditions were within the confidence interval for all models in the case of LR and W tests. When the generating model is *A*-CDM, the LM test tended to be conservative (i.e., Type I error dropped close to 0).

None of the other interactions for the LR, W, and LM tests were relevant, so the main effects could be interpreted. However, as noted above, Type I error was generally acceptable only in the high item quality condition. Sample size and test length affected the performance of the three statistics: Sample size had a small effect for the LR, W, and LM tests ($\hat{\omega}^2$ = .01, .03, and .01, respectively), whereas test length had a small effect on the Type I error of the LR and LM tests ($\hat{\omega}^2$ = .02 and .03, respectively), and a large effect in the case of W test ($\hat{\omega}^2$ = .17). The Type I error was closer to the nominal level as the sample size and the test length increased. As can be observed in Online Annex 2, there were cases where Type I error was within the confidence interval when the test length and the sample size were large (i.e., J = 24 or 36 and N = 1,000). Finally, correlational structure had a small effect in the case of the LM test ($\hat{\omega}^2$ = .02). The Type I error for the LM test was inflated in the bidimensional conditions compared with the unidimensional conditions, although differences were small.

Power

The $\hat{\omega}^2$ values and marginal means associated with each main effect on the power are provided in Table 2. For most of the conditions involving high-quality items, it was not necessary to correct α_a. For example, α_a was corrected for the LR tests only in some of the conditions (i.e., J = 12 and N = 500). The pattern of effects of the manipulated factors on the power was very similar for all the tests. However, power of the LR and W tests was almost always better than those of the $S - X^2$ and LM tests—the grand means across models were 0.75, 0.78, 0.25, and 0.46 for LR, W, $S - X^2$, and LM tests, respectively. Again, item quality had the greatest effect with an average $\hat{\omega}^2$ = .74. Power was usually lower than .80 in the low item quality conditions for all the statistics. This factor was involved in all the salient high-order interactions: Sample

Table 2. Marginal Means and Effect Sizes of the ANOVA Main Effects for the Power for Rejecting a False Reduced Model.

		Data factor/level																	
		N			DIM			*J*				IQ				Generating, true model (MOD)			
Fitted, false model	Item-fit statistic	$\hat{\omega}^2$	500	1,000	$\hat{\omega}^2$	Uni	Bi	$\hat{\omega}^2$	12	24	36	$\hat{\omega}^2$	LD	MD	HD	$\hat{\omega}^2$	A-CDM	DINO	Grand mean
DINA	$S-X^2$	**.25**	.16	.25	.00	.21	.20	**.40**	.12	.22	.27	**.81**	.07	.13	.42	**.53**	.13	.29	0.21
	LR	**.13**	.68	.79	**.03**	.76	.71	**.26**	.62	.76	.82	**.78**	.35	.85	1.00	**.18**	.67	.80	0.73
	W	**.14**	.74	.84	**.07**	.82	.75	**.22**	.70	.81	.86	**.76**	.48	.89	1.00	**.22**	.72	.86	0.79
	LM	**.02**	.66	.70	.00	.67	.69	**.13**	.63	.67	.74	**.62**	.53	.61	.90	**.82**	.95	.42	0.68
																	DINA	DINO	
A-CDM	$S-X^2$	**.29**	.23	.35	**.03**	.28	.31	**.15**	.24	.30	.34	**.86**	.09	.16	.63	**.20**	.34	.24	0.29
	LR	**.18**	.64	.75	.00	.69	.69	**.39**	.56	.72	.80	**.89**	.22	.85	1.00	**.09**	.73	.65	0.69
	W	**.22**	.65	.77	.00	.71	.71	**.37**	.60	.72	.81	**.89**	.27	.87	1.00	**.04**	.73	.69	0.71
	LM	**.04**	.09	.13	**.15**	.07	.15	**.16**	.05	.13	.15	**.59**	.05	.02	.28	**.05**	.09	.13	0.11
																	DINA	A-CDM	
DINO	$S-X^2$	**.21**	.22	.31	**.03**	.25	.28	**.43**	.17	.27	.35	**.86**	.07	.17	.55	**.58**	.37	.16	0.26
	LR	**.10**	.78	.86	.01	.81	.83	**.32**	.71	.85	.91	**.76**	.51	.96	1.00	**.34**	.91	.73	0.82
	W	**.13**	.80	.88	**.01**	.83	.85	**.38**	.73	.87	.92	**.79**	.56	.97	1.00	**.39**	.92	.76	0.84
	LM	**.06**	.57	.63	.00	.60	.60	**.01**	.58	.61	.62	**.28**	.51	.61	.69	**.90**	.25	.96	0.60

Note. Effect size values greater than .010 are shown in bold. Shaded cells correspond to power in the [0.80, 1.00] interval. *N* = sample size; DIM = correlational structure; *J* = test length; IQ = item quality; MOD = generating model; Uni = unidimensional; Bi = bidimensional; LD = low discrimination; MD = medium discrimination; HD = high discrimination; DINA = deterministic inputs, noisy and gate; CDM = cognitive diagnosis model; *A*-CDM = additive CDM; DINO = deterministic inputs, noisy "or" gate.

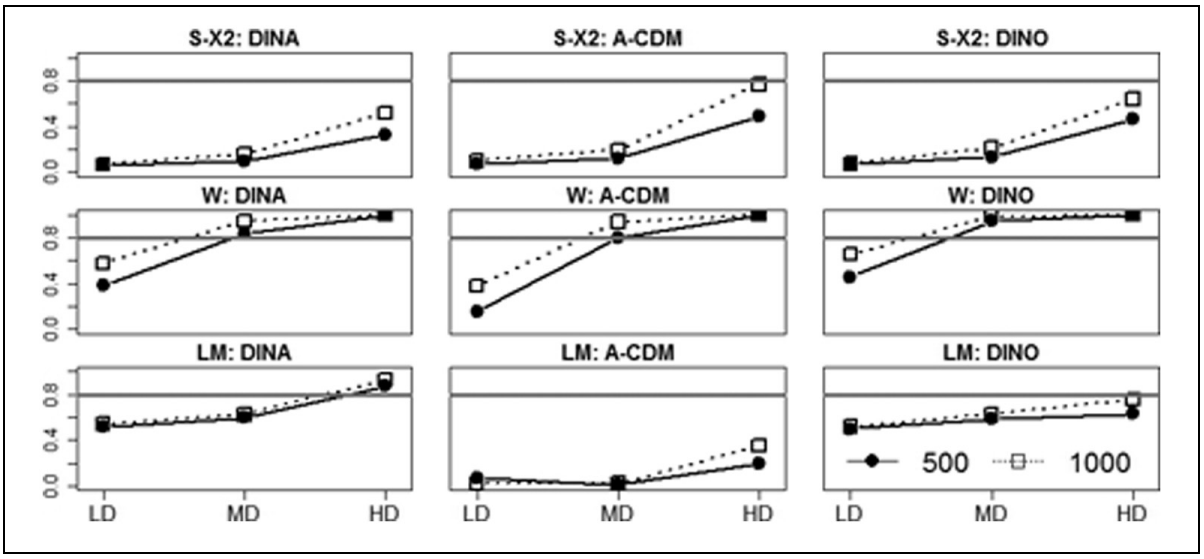

Figure 2. Two-way interaction of Sample Size × Item Quality with $S-X^2$, W, and LM power as dependent variables.

Note. The horizontal gray line represents a statistical power of .80. DINA = deterministic inputs, noisy and gate; CDM = cognitive diagnosis model; *A*-CDM = additive CDM; DINO = deterministic inputs, noisy "or" gate.

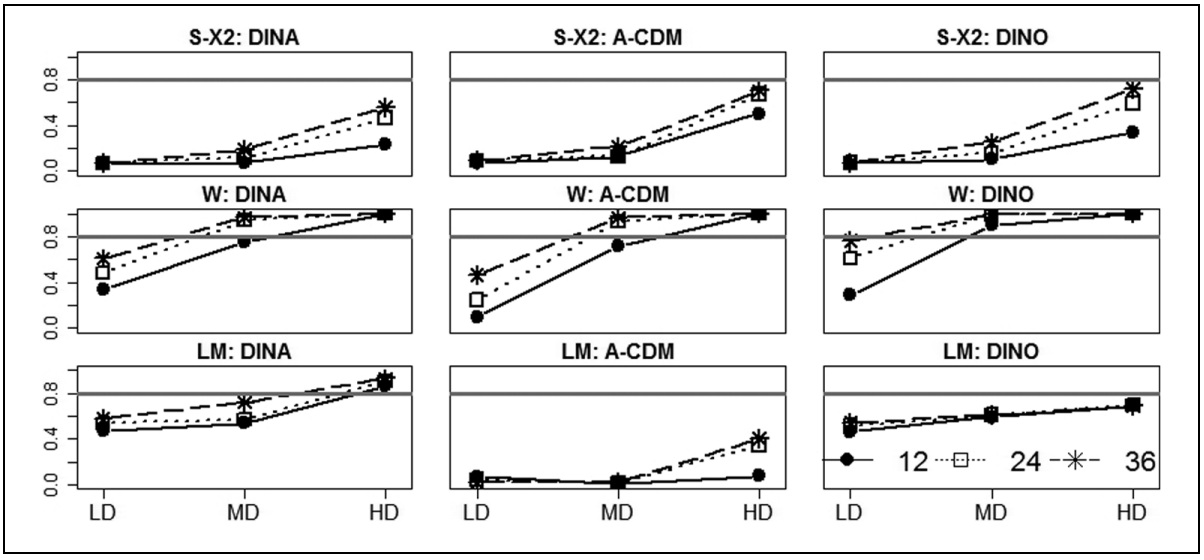

Figure 3. Two-way interaction of Test Length × Item Quality with $S-X^2$, LR, and LM power as dependent variables.

Note. The horizontal gray line represents a statistical power of .80. DINA = deterministic inputs, noisy and gate; CDM = cognitive diagnosis model; *A*-CDM = additive CDM; DINO = deterministic inputs, noisy "or" gate.

Size × Item Quality (Figure 2), Test Length × Item Quality (Figure 3), Test Length × Item Quality × Correlational Structure (Figure 4), and Test Length × Item Quality × Generating Model (Figure 5). Here follows a description of each of these interactions.

As noted before, power increased as the item quality got better. This effect interacted with the sample size and test length (see Figures 2 and 3). In the case of the $S-X^2$ and LM tests, the improvement on the power associated with moving from low- to medium-quality items was similar for the different levels of sample size and test length, but this gain is generally much bigger

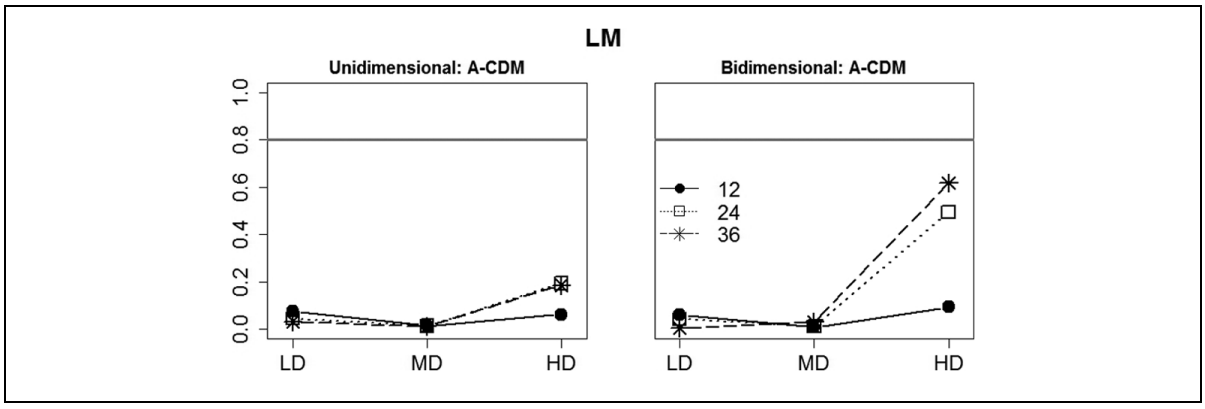

Figure 4. Three-way interaction of Correlational Structure × Item Quality × Test Length with LM power for rejecting *A*-CDM when it is false as dependent variable.
Note. The horizontal gray line represents a statistical power. CDM = cognitive diagnosis model; *A*-CDM = additive CDM.

when moved from medium- to high-quality items in the case of the N = 1,000, J = 24, and J = 36 conditions. The pattern of results for the LR test was similar to the one observed for the W test. Thus, only the W test is depicted in Figure 3. Power in the medium-quality item conditions was already close to 1.00 when N = 1,000 and J = 24 or 36. This is why there is a small room for improvement when moved to high-quality item conditions because of this ceiling effect.

In the case of the LM test, it was found that the three-way Correlational Structure × Test Length × Item Quality had a salient effect on the power for rejecting *A*-CDM when it was false. As can be seen from Figure 4, only test length and item quality had a noteworthy effect on the LM power in the bidimensional scenario.

There is a salient interaction effect of the item quality and the generating model factors affecting all the statistics. As can be observed from Table 2, in general, the main effect of the generating model indicates that, for $S - X^2$, LR, and W tests, the DINA model was easier to reject when the data were generated with the DINO model and vice versa. Power for rejecting *A*-CDM was generally higher when data were generated with the DINA model. The effect on the power of LM was different: The power for rejecting DINA and DINO models was higher for data generated using the *A*-CDM, and the power for rejecting *A*-CDM was close to 0, regardless the generating model − .09 and .13 for data generated with DINA and DINO models, respectively. In short, LM tended to reject models different from *A*-CDM. In the case of the $S - X^2$ power, power increased as the item quality got better, but the increment was larger for models which were easier to distinguish (i.e., DINA vs. DINO, *A*-CDM vs. DINA). This relationship between item quality and generating models was affected by the test length in the case of LR, W, and LM tests. This three-way interaction was very similar for the LR and W tests, so it was only depicted for the W test (see Figure 5). Power was always equal to 1.00 in the high item quality conditions, regardless of the test length. In the medium item quality conditions, power was also very high when comparing the more distinguishable models (i.e., DINA vs. DINO, *A*-CDM vs. DINA), even when test was composed of a small number of items (J = 12). In the low item quality conditions, the LR and W tests only can differentiate between the DINA and DINO models, but only if the number of items was at least 24. In the case of the LM test, this three-way interaction had only a salient effect on the power for rejecting DINA and *A*-CDM models. However, power was generally only acceptable for rejecting DINA and DINO models when the generating model is *A*-CDM, regardless of the test length and the quality of the items.

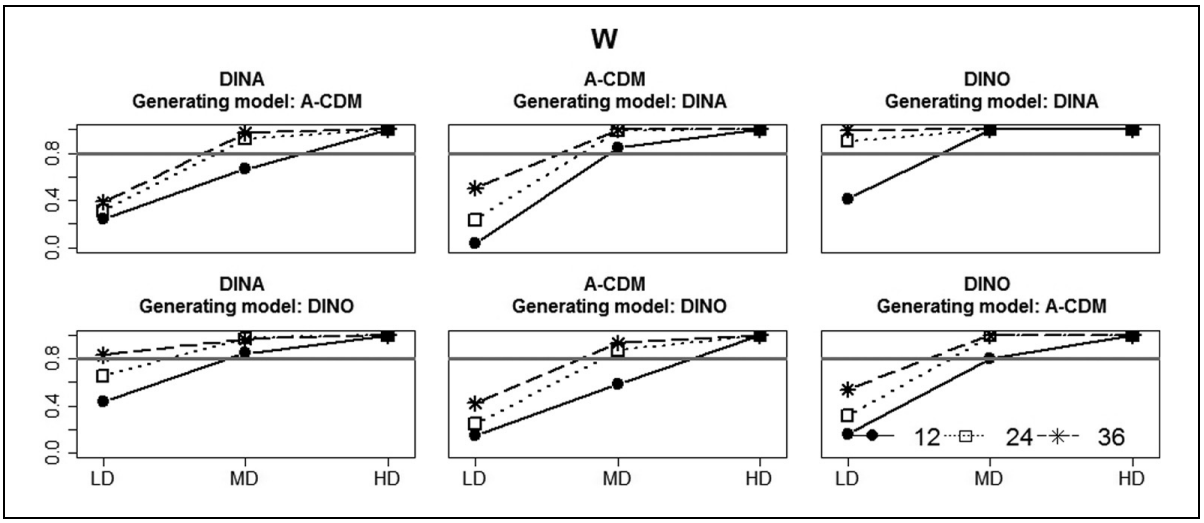

Figure 5. Three-way interaction of Generating Model × Item Quality × Test Length for W test power for rejecting DINA, *A*-CDM, and DINO when they are false as dependent variables.

Note. The horizontal gray line represents a statistical power of .80. DINA = deterministic inputs, noisy and gate; CDM = cognitive diagnosis model; *A*-CDM = additive CDM; DINO = deterministic inputs, noisy "or" gate.

Discussion

Even though the interest in CDMs began in response to the growing demand for a better understanding of what students can and cannot do, CDMs have been recently applied to data from different contexts such as psychological disorders (de la Torre et al., 2015; Templin & Henson, 2006) and competency modeling (Sorrel et al., 2016). Item quality has been found to be typically low outside of the educational context. In addition, according to the literature this is an expected result of applications where the attributes are specified post hoc (i.e., CDMs are retrofitted; Rupp & Templin, 2008). The suitable application of a statistical model requires the assessment of model–data fit. One important question that is raised by these new applications is how item quality may affect the available procedures for assessing model fit. While extensive studies have been conducted to evaluate the performance of various fit statistics at the test (e.g., Chen et al., 2013; Y. Liu et al., 2016) and person levels (e.g., Cui & Leighton, 2009; Y. Liu et al., 2009), the item level is probably the one that has received less attention in the previous literature. The statistical properties of the item-fit statistics remain unknown (e.g., $S-X^2$, LR, and LM tests) or need further investigation (e.g., W test). Taking the above into account, this study provides information about the usefulness of these indexes on different plausible scenarios.

To use item-fit statistics in practical use, it is necessary that Type I error is close to the nominal value, and that they have a great power to reject false models. In the case of the statistic evaluating absolute fit, $S-X^2$, although it has been found to have a satisfactory Type I error, its power is far from reaching acceptable values. These results are in line with previous studies assessing the performance of χ^2-like statistics in the context of the DINA model (C. Wang et al., 2015). Here, these results are extended to compensatory and additive models (i.e., DINO and *A*-CDM). In conclusion, given its poor performance in terms of power, decisions cannot be made based only on this indicator. There are, however, a number of possible solutions for dealing with this problem that need to be considered in future studies. For example, C. Wang et al. (2015) have shown how Stone's (2000) method can be applied to avoid low power in the case of the DINA model. To the authors' knowledge, this method has not yet been included in the software available.

Overall, the Type I error and power comparisons favor LR and W tests over the LM test. However, and more importantly, Type I error is only acceptable (i.e., $\alpha \cong .05$) when the item quality is high: With a very few exceptions, Type I error with medium- and low-quality items is generally inflated. These results are tentatively attributed to the noise in the estimation of the item parameters and the standard errors in those conditions. This also applies to other contexts such as the evaluation of DIF (e.g., Bai, Sun, Iaconangelo, & de la Torre, 2016). Particularly in the case of the LR test, in medium item quality conditions this can be compensated by an increase in the number of respondents and items when the true model is DINA or *A*-CDM. For the DINO model, Type I error is highly inflated even in those conditions, which is not consistent with the previous results of de la Torre and Lee (2013). However, when the actual alpha is corrected, so that it corresponds to the nominal level, it was found that the power is still generally high in the medium item quality conditions. Monte Carlo methods can be used in practical settings to approximate the distribution of the statistics under the null hypothesis as it is done in the simulation study (e.g., Rizopoulos, 2006). All things considered, this means that, most likely, an incorrect model will not be chosen if LR or W test is used and the item quality is at least medium, which is consistent with de la Torre and Lee's results for the W test. However, this does not mean that CDMs cannot be applied in poor-quality items conditions. In these situations, the model fit of the test should be assessed as a whole, and it should be ensured that the derived attribute scores are valid and reliable. Another promising alternative is to use a strategy that makes the best of each statistic. According to the results of the present study, $S-X^2$, LR, and W statistics can be used simultaneously as a useful tool for assessing item fit in empirical applications. Among all the models fitting the data according to the $S-X^2$ statistic, the one pointed by the LR or the W test will be chosen as the most appropriate model.

Even though the LR test was found to be relatively robust than the W test, the power of W test was slightly higher. Another advantage of using the W test is that it requires only the unrestricted model to be estimated. In contrast, the LR test required $J_{K_j^*>1}NR+1$ models to be estimated, where *NR* is the number of reduced models to be tested. For example, for one of the conditions with 36 items and 1,000 examinees the computation of the LR and W tests requires 2.44 min and 6 s, respectively. In other words, the W test was 24 times fast than the LR test. Furthermore, in a real scenario, multiple CDMs can be fitted within the same test. Thus, a more exhaustive application of the LR test would require comparing the different combinations of the models, and lead to substantially longer time to implement the LR test. Future studies should explore how this limitation can be addressed.

Although the LM test was introduced as an alternative for assessing fit at the item level, it was found that its performance is highly affected by the underlying model: It tended to keep *A*-CDM and reject DINA and DINO models. This test focuses on the distance between the restricted and the unrestricted item parameter estimates. A possible explanation for this poor performance is that the computation of this difference (i.e., the score function) relies on a good estimation of the attribute joint distribution. In this regard, Rojas et al. (2012) found that fitting an incorrect reduced CDM may have a great impact on the attribute classification accuracy, affecting the estimation of the attribute joint distribution, and thus the performance of this test.

To fully appreciate the current findings, some caveats are in order. A first caveat relates to the number of attributes. In certain application fields, the number of attributes can be high. For example, Templin and Henson (2006) specify 10 attributes corresponding to the 10 *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text rev.; DSM-IV-TR; American Psychiatric Association [APA], 2000) criteria for pathological gambling. Thus, it is recommended that future research examines the effect of the number of attributes. Second, all items were simulated to have the same discrimination power. In a more realistic scenario, discriminating and nondiscriminating items are mixed. Third, the present study focuses on inferential

statistical evaluation. Future studies should consider other approximations. For example, goodness-of-fit descriptive measures have been shown to be useful in some situations. Chen et al. (2013) found that fit measures based on the residuals can be effectively used at the test level. Kunina-Habenicht, Rupp, and Wilhelm (2012) found that the distributions of the root mean square error of approximation (RMSEA) and median absolute deviation (MAD) indexes can be insightful when evaluating models and Q-matrices in the context of the log-linear model framework. New studies might try to extend this result to other general frameworks.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by Grant PSI2013-44300-P (Ministerio de Economia y Competitividad and European Social Fund).

Supplemental Material

The online supplementary appendices are available at http://journals.sagepub.com/doi/suppl/10.1177/0146621617707510.

References

Bai, Y., Sun, Y., Iaconangelo, C., & de la Torre, J. (2016, July). *Improving the Wald test DIF detection under CDM framework*. Paper presented at the International Meeting of Psychometric Society, Asheville, NC.

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange Multiplier tests: An expository note. *American Statistician*, *36*, 153-157.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*, 123-140.

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement*, *46*, 429-449.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179-199.

de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*, 355-373.

de la Torre, J., van der Ark, L. A., & Rossi, G. (2015). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*. Advance online publishing. doi:10.1177/0748175615569110

DiBello, L., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. V. Rao & S. Sinharay (Eds.), *Handbook of Statistics: Vol. 26. Psychometrics* (pp. 979-1027). Amsterdam, The Netherlands: Elsevier.

Engle, R. F. (1983). Wald, likelihood ratio, and lagrange multiplier tests in econometrics. In M. D. Intriligator & Z. Griliches (Eds.), *Handbook of econometrics* (Vol. *2*., pp. 796-801). New York, NY: Elsevier.

Glas, C. A. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, *64*, 273-294.

Glas, C. A., & Suárez-Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*, 87-106.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 333-352.

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois, Urbana-Champaign.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191-210.

Kang, T., & Chen, T. T. (2008). Performance of the Generalized S-X2 Item Fit Index for Polytomous IRT Models. *Journal of Educational Measurement*, *45*, 391-406.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*, 59-81.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, *33*, 519-537.

Liu, H.-Y., You, X.-F., Wang, W.-Y., Ding, S.-L., & Chang, H.-H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification*, *30*, 152-172.

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement*, *33*, 579-598.

Liu, Y., Tian, W., & Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics*, *41*, 3-26.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score ''equatings.'' *Applied Psychological Measurement*, *8*, 453-461.

Ma, W., & de la Torre, J. (2016). GDINA: The generalized DINA model framework (R package Version 0.9.8). Retrieved from http://CRAN.R-project.org/package=GDINA

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*, 200-217.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187-212.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50-64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*, 289-298.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, *34*, 782-799.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response analysis. *Journal of Statistical Software*, *17*, 1-25.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2015). CDM: Cognitive Diagnosis Modeling (R package Version 4. 6–0). Retrieved from http://CRAN.R-project.org/package=CDM

Rojas, G., de la Torre, J., & Olea, J. (2012, April). *Choosing between general and specific cognitive diagnosis models when the sample size is small*. Paper presented at the meeting of the National Council on Measurement in Education, Vancouver, Canada.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*, 219-262.

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitively diagnostic models—A case study. *Educational and Psychological Measurement*, *67*, 239-257.

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*, *19*, 506-532.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness of fit test statistic in IRT models. *Journal of Educational Measurement*, *37*, 58-75.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*, 331-352.

Tatsuoka, K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453-488). Hillsdale, MI: Lawrence Erlbaum.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287-305.

von Davier, M. (2005). A General diagnostic model applied to language testing data (ETS research report). Princeton, NJ: Educational Testing Service.

Wang, C., Shu, Z., Shang, Z., & Xu, G. (2015). Assessing item level fit for the DINA model. *Applied Psychological Measurement*, *39*, 525-538.

Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*, 479-498.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245-262.

Chapter 4

Two-Step Likelihood Ratio Test for Item-Level Model Comparison in Cognitive Diagnosis Models

This study was published in the journal *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* (Sorrel, de la Torre, Abad, & Olea, 2017). The following pages include this publication. This article was first published on June 2, 2017. The issue has not been published yet. The article can be downloaded from https://doi.org/10.1027/1614-2241/a000131. The journal impact factor and 5-year impact factor (2016) are 1.143 and 1.632, respectively. I was awarded with the *Young Methodologist 2016 European Association of Methodology award* at the conference held in Mallorca (Spain) where I first presented this study.

Original Article

Two-Step Likelihood Ratio Test for Item-Level Model Comparison in Cognitive Diagnosis Models

Miguel A. Sorrel,[1] Jimmy de la Torre,[2] Francisco J. Abad,[1] and Julio Olea[1]

[1]Department of Social Psychology and Methodology, Universidad Autónoma de Madrid, Spain
[2]Faculty of Education, The University of Hong Kong, Hong Kong

Abstract: There has been an increase of interest in psychometric models referred to as cognitive diagnosis models (CDMs). A critical concern is in selecting the most appropriate model at the item level. Several tests for model comparison have been employed, which include the likelihood ratio (LR) and the Wald (W) tests. Although the LR test is relatively more robust than the W test, the current implementation of the LR test is very time consuming, given that it requires calibrating many different models and comparing them to the general model. In this article, we introduce the two-step LR test (2LR), an approximation to the LR test based on a two-step estimation procedure under the *generalized deterministic inputs, noisy, "and" gate* (G-DINA) model framework, the two-step LR test (2LR). The 2LR test is shown to have similar performance as the LR test. This approximation only requires calibration of the more general model, so that this statistic may be easily applied in empirical research.

Keywords: cognitive diagnosis models, model comparison, item fit, Type I error, power

Cognitive diagnosis models (CDMs) have received increasing attention within the field of educational and psychological measurement. These models are useful tools to provide diagnostic information about examinees' cognitive profiles in domains such as education (e.g., Lee, Park, & Taylan, 2011), measurement of psychological disorders (e.g., de la Torre, van der Ark, & Rossi, 2015), and competency modeling (e.g., Sorrel, et al., 2016). Selection of an appropriate CDM is based in part on model-data fit. Model-data fit can be assessed at the test level (e.g., Chen, de la Torre, & Zhang, 2013; Liu, Tian, & Xin, 2016). If the model, particularly if it has a general formulation, fits the data, then it may be useful to study hypothesis about differences in response processes across items. Relative fit indices can be used to evaluate the discrepancy among different statistical models. According to a recent evaluation on the performance of various goodness-of-fit statistics for relative fit evaluation at the item level, the likelihood ratio (LR) test is more robust than other statistics (Sorrel, Abad, Olea, Barrada, & de la Torre, 2017).

The current implementation of the LR test is very time consuming, given that it requires to calibrate many different models and compare them to the general model. For this reason, the Wald (W) test (de la Torre & Lee, 2013) is generally preferred. In light of this, the primary purpose of this study is to investigate the performance of an approximation to the LR test, the two-step LR (2LR) test, which only requires to estimate the more general model once. Reduced model item parameters are estimated at the item level (i.e., one item at a time) rather than at the test level (i.e., all the items in the test simultaneously) following an alternative, heuristic estimation procedure originally introduced by de la Torre and Chen (2011) and further explored here. This procedure is based on the the *generalized deterministic inputs, noisy, "and" gate* (G-DINA; de la Torre, 2011) model framework. The rest of the article is structured as follows. Next section provides background information about CDMs and model comparison. The design of the simulation study is described thereafter. Subsequently, some results are presented to demonstrate the performance of the estimation procedure and the performance of the new statistic compared to the LR and W tests. The last section provides the concluding remarks.

Cognitive Diagnosis Modeling

Cognitive Diagnosis Modelings are multidimensional, categorical latent-trait models developed primarily for identifying which attributes (e.g., skills, mental disorders, competencies)

Methodology (2017), 13(Supplement), 39–47
DOI: 10.1027/1614-2241/a000131

are mastered and which ones do not (see, e.g., Rupp & Templin, 2008, for an overview of these models). For an assessment diagnosing K attributes, examinees are grouped into 2^K latent classes. Latent classes are represented by an attribute vector denoted by $\boldsymbol{\alpha}_l = (\alpha_{l1}, \alpha_{l2}, \ldots, \alpha_{lK})$, where $l = 1, \ldots, 2^K$. Specifically, $\alpha_{lK} = 1$ or 0 represents mastery or nonmastery of attribute k, respectively. In each latent class, examinees all have the same probability of success on a particular item j, denoted by $P(X_j = 1|\boldsymbol{\alpha}_l) = P_j(\boldsymbol{\alpha}_l)$. In other contexts (e.g., measurement of psychological disorders), $P_j(\boldsymbol{\alpha}_l)$ indicates the probability of item endorsement. The attributes that are required to correctly answer each item are defined in a $J \times K$ matrix, commonly known as Q-matrix (Tatsuoka, 1990), where J is the test length.

Several general models that encompass reduced CDMs have been proposed, including the above-mentioned G-DINA model. The G-DINA model is a generalization of the *deterministic inputs, noisy, "and" gate* (DINA; Haertel, 1989) model that describes the probability of success on item j in terms of the sum of the effects of the attributes involved and their corresponding interactions. Let the number of required items for item j be denoted by K_j^*. In this model, latent classes are sorted into $2^{K_j^*}$ latent groups. Each of these latent groups represents one reduced attribute vector $\boldsymbol{\alpha}_{lj}^*$. The probability of success associated to $\boldsymbol{\alpha}_{lj}^*$ is defined as

$$P(\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}, \quad (1)$$

where δ_{j0} is the intercept or baseline probability for item j, δ_{jk} is the main effect due to α_k, $\delta_{jkk'}$ is the interaction effect due to α_k and $\alpha_{k'}$, and $\delta_{j12\ldots K_j^*}$ is the interaction effect due to $\alpha_1,\ldots,\alpha_{K_j^*}$. Thus, there are $2^{K_j^*}$ parameters to be estimated for item j.

By constraining the parameters of the saturated model, de la Torre (2011) has shown that some of the commonly used reduced CDMs can be obtained, including the DINA model and the *additive* CDM (*A*-CDM; de la Torre, 2011). To compare the different models in a more straightforward manner, this article uses $\boldsymbol{\varphi}_j$ to represent reduced model item parameters across all reduced CDMs. Namely, φ_{j0} is the intercept for item j, φ_{jk} is the main effect due to α_k, and $\varphi_{j12\cdots K_j^*}$ is the interaction effect due to $\alpha_1, \ldots, \alpha_{K_j^*}$. The DINA model is a conjunctive model, that is, an examinee needs to have mastered all required attributes to correctly answer a particular item. As such, the DINA model separates examinees into two latent groups for each item: one group with examinees who have mastered all attributes required by the item and one group with examinees lacking at least one. The probability of correct response is represented by the DINA model as follows:

$$P(\boldsymbol{\alpha}_{lj}^*) = \varphi_{j0} + \varphi_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ljk}. \quad (2)$$

Therefore, the DINA model has two parameters per item and is deduced from the G-DINA model by setting to zero all terms except for δ_{j0} and $\delta_{j12\ldots K_j^*}$ to zero. For the *A*-CDM, all the interaction terms are dropped. The item response function is given by

$$P(\boldsymbol{\alpha}_{lj}^*) = \varphi_{j0} + \sum_{k=1}^{K_j^*} \varphi_{jk}\alpha_{jlk}. \quad (3)$$

This is the G-DINA without the interaction terms, and it shows that mastering attribute α_{lk} raises the probability of success on item j by φ_{jk}. There are $K_j^* + 1$ parameters for item j in the *A*-CDM. In this respect, the DINA model involves a conjunctive process, whereas the *A*-CDM involves an additive process. Figure 1 gives a graphical representation of an item requiring two attributes when it conforms to the DINA, *A*-CDM, or the G-DINA model. As can be observed from Figure 1, in the DINA model latent classes are sorted into two latent groups. Examinees who have mastered all attributes required by the item have a probability of correct response equal to $\varphi_{j0} + \varphi_{j12\cdots K_j^*}$. Examinees lacking at least one attribute will have a probability of correct response equal to the baseline probability (i.e., φ_{j0}). In the case of the *A*-CDM, each attribute has a main impact. For example, examinees mastering only the first attribute will have a probability of success equal to $\varphi_{j0} + \varphi_{j1}$.

Model Comparison in CDM

Each CDM assumes a different cognitive process involved in responding to an item (e.g., conjunctive or additive). The task in model selection is to select the model that is the best fit to the data. For nested CDMs, model selection at the item level can be done using the three common tests for assessing relative fit (Buse, 1982): likelihood ratio (LR), Wald (W), and Lagrange multiplier (LM) tests. In all the three cases, the statistic is assumed to be asymptotically χ^2 distributed with $2^{K_j^*} - p$ degrees of freedom, where p is the number of parameters of the reduced model.

To investigate the finite sample performance of these tests, Sorrel et al. (2017) conducted a simulation study. Overall the Type I error and power comparisons favored LR and W tests over the LM test. LR was found to be relatively more robust than the W test. However, the appealing advantage of using the W test is that it required only the

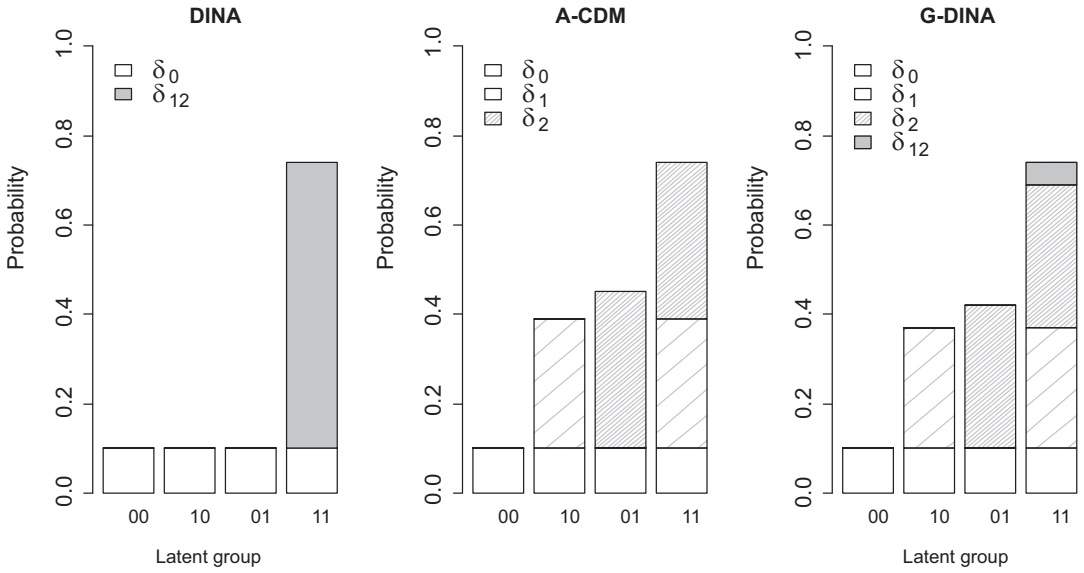

Figure 1. This figure depicts the probability of correctly answering an item requiring two attributes for the DINA, A-CDM, and G-DINA models. Item parameters are denoted by δ.

unrestricted model (i.e., G-DINA) to be estimated. In contrast, the LR test required $J^* \cdot \mathrm{NR} + 1$ models to be estimated, where J^* is the number of items measuring more than one attribute and NR is the number of reduced models to be tested. In this study, we propose an approximation to the LR test, 2LR, which also has the appealing advantage of only requiring the G-DINA model to be estimated. In the following, we will describe how the 2LR test is computed.

Approximation to the LR test

The LR test is a statistical test used to compare the goodness-of-fit of two models, one of which is nested in the other. Because adding additional parameters to a more general model will always result in a higher likelihood, CDMs with general formulations will provide a better fit to the data. The LR test provides one objective criterion for evaluating if the more general model fits a particular dataset significantly better. In the traditional implementation of the LR test in the CDM context, the more general model, the G-DINA model, is estimated for all the items. This model is compared with a reduced model fitted to a target item, whereas the G-DINA model is fitted to the rest of the items. Both model specifications are estimated and the LR statistic is computed as twice the difference in the log-likelihoods. The application of the LR test requires comparing the different combinations of the models. To obtain the likelihood of a model, both item parameter and posterior distribution estimates are needed. Rojas, de la Torre, and Olea (2012) found that the attribute classification accuracy of the G-DINA model is the best when the underlying model is not known. de la Torre and Chen (2011) introduced a procedure for estimating the reduced model item parameters using the attribute classification obtained with the G-DINA. Let us review their proposal.

Two-Step Estimation Procedure

de la Torre and Chen (2011) originally introduced an alternative estimation procedure that uses the G-DINA estimates for efficiently estimating the parameters of several reduced CDMs. This method is referred to as two-step estimation procedure because the estimation of the item parameters (i.e., $\boldsymbol{\varphi}_j$) for the reduced CDMs is done in two steps. The first step involves estimating the G-DINA model parameters, $P_j = \left\{\boldsymbol{P}\left(\boldsymbol{\alpha}_{lj}^*\right)\right\}$. The second step involves computing the corresponding $\boldsymbol{\varphi}_j$ of the reduced models. de la Torre (2011) showed that $\boldsymbol{P}_j$ can be estimated using an expectation-maximization (EM) implementation of the marginal maximum likelihood (MML) estimation. Briefly, it can be shown that the MML estimate of the parameter $P\left(\boldsymbol{\alpha}_{lj}^*\right)$ is given by

$$\hat{P}\left(\boldsymbol{\alpha}_{lj}^*\right) = \frac{R_{\alpha_{lj}^*}}{I_{\alpha_{lj}^*}}, \tag{4}$$

where $I_{\alpha_{lj}^*}$ and $R_{\alpha_{lj}^*}$ are the expected number of examinees and correct responses in the latent group $\boldsymbol{\alpha}_{lj}^*$, respectively.

Once $\mathbf{P}_j$ has been estimated, item parameters φ_j can be obtained through some linear transformations or maximization processes. For DINA model, a $2^{K_j^*} \times p$ design matrix $\mathbf{M}$

can be used to linearly transform the G-DINA model parameters into reduced model item parameters, where p is the number of model parameters. To illustrate, let $K_j^* = 2$. The saturated design matrix is

$$\mathbf{M}_{4\times 4}^{(\mathrm{S})} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}. \tag{5}$$

In deriving the reduced models, subsets of functions of subsets of the columns of $\mathbf{M}^{(\mathrm{S})}$ are used. For example, the design matrix for the DINA model would be

$$\mathbf{M}_{4\times 2} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}. \tag{6}$$

The design matrix for the DINA model indicates that all elements of $\mathbf{P}_j$ contain φ_{j0}, whereas only the last element contains $\varphi_{j12\cdots K_j^*}$. Several elements of $\mathbf{P}_j$ need to be combined to obtain $\boldsymbol{\varphi}_j$. These elements are differentially weighted to account for the relative size of the latent classes. DINA model estimates are obtained by

$$\hat{\boldsymbol{\varphi}}_j = (\mathbf{M}'\mathbf{W}\mathbf{M})^{-1}\mathbf{M}'\mathbf{W}\hat{\mathbf{P}}_j, \tag{7}$$

where $\mathbf{W}$ is a diagonal matrix $\mathbf{W}_{2^{K_j^*}\times 2^{K_j^*}} = \{I_{\boldsymbol{\alpha}_{lj}^*}\}$ and $\hat{\mathbf{P}}_j = \{\hat{P}(\boldsymbol{\alpha}_{lj}^*)\}$.

For *A*-CDM, however, the design matrix cannot be used because $\boldsymbol{\varphi}_j$ cannot be expressed as a simple linear combination of the elements of $\mathbf{P}_j$. Instead, the parameter estimates can be obtained by maximizing the likelihood of $\boldsymbol{\varphi}_j$ given $\mathbf{R}_j = \{\mathrm{R}_{\boldsymbol{\alpha}_{lj}^*}\}$ and $\mathbf{I}_j = \{\mathrm{I}_{\boldsymbol{\alpha}_{lj}^*}\}$ obtained in the first step as follows:

$$\mathbf{L}(\boldsymbol{\varphi}_j|\mathbf{R}_j, \mathbf{I}_j) = \prod_{l=1}^{2^{K_j^*}} P^{(R)}(\boldsymbol{\alpha}_{lj}^*)^{R_{\boldsymbol{\alpha}_{lj}^*}} \left[1 - P^{(R)}(\boldsymbol{\alpha}_{lj}^*)\right]^{\left(I_{\boldsymbol{\alpha}_{lj}^*} - R_{\boldsymbol{\alpha}_{lj}^*}\right)}, \tag{8}$$

where $P^{(R)}(\boldsymbol{\alpha}_{lj}^*)$ is the probability of success implied by the reduced model. In this article we explore how this estimation procedure can be used as a basis in efficiently computing an approximation to the LR test.

Two-Step Likelihood Ratio Test

Item-level maximum likelihoods for the saturated and reduced models can be computed based on the estimated item parameters and attribute distribution. Item parameters are those estimated for G-DINA and the reduced model (i.e., DINA or *A*-CDM), whereas the attribute distribution is obtained in the first step based on the G-DINA model. Comparing the two marginalized likelihoods using a LR test can be useful to find out if a reduced model is appropriate for those items measuring more than one attribute. We proposed the 2LR test as an efficient way of computing the statistic, and it is computed as

$$\begin{aligned} 2\mathrm{LR}_j &= 2\left[\log L(\mathbf{P}_j|\mathbf{R}_j, \mathbf{I}_j) - \log L\left(\boldsymbol{\varphi}_j|\mathbf{R}_j, \mathbf{I}_j\right)\right] \\ &\sim \chi^2\left(2^{K_j^*} - p\right), \end{aligned} \tag{9}$$

where $\mathbf{P}_j$ is the vector of GDINA item parameters and $\boldsymbol{\varphi}_j$ is the vector of reduced model parameters for item j. The likelihood function that is employed is the one represented in (8). Compared to the LR test, only one model (i.e., G-DINA) is estimated. Given that the two-step estimation procedure is the basis of the new statistic, it is pivotal to ensure its accuracy under plausible scenarios.

Method

A simulation study was conducted to assess the accuracy of the two-step item parameter estimates and performance of the 2LR test compared to the LR and W tests. Four factors were varied and their levels were chosen to represent realistic scenarios. These factors were: (1) generating model (*MOD*; DINA model and *A*-CDM); (2) test length (*J*; 30 and 60 items); (3) sample size (*N*; 500, 1,000, and 2,000 examinees); and (4) item quality or discrimination, defined as the difference between the maximum and the minimum probabilities of correct response according to the attribute latent profile (IQ; .40, .60, and .80).

The probabilities of success for individuals who mastered none of the required attributes were fixed to .30, .20, and .10 for the low, medium, and high item quality conditions, respectively; the corresponding probabilities for those who mastered all of the required attributes were fixed to .70, .80, and .90. For the *A*-CDM, an increment of $.40/K_j^*$, $.60/K_j^*$, and $.80/K_j^*$ was associated with each attribute mastery for the low, medium, and high item quality conditions, respectively. The number of attributes was fixed to $K = 5$. The Q-matrix used in simulating the response data and fitting the models is given in Table 1. This Q-matrix was constructed such that each attribute appears alone, in a pair, or in a triple the same number of times as other attributes. For $J = 60$, each item was used twice.

For each of the 36 factor combinations, 200 datasets were generated and DINA, *A*-CDM, and G-DINA models were fitted. We evaluated whether the two-step algorithm is comparable, in terms of estimation accuracy or variability, to the standard EM-MML algorithm. For comparison of estimation accuracy, we computed the bias, $\hat{\varphi} - \varphi$; for

Table 1. Simulation study Q-matrix for the $J = 30$ conditions

	Attribute				
Item	α_1	α_2	α_3	α_4	α_5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
6	1	0	0	0	0
7	0	1	0	0	0
8	0	0	1	0	0
9	0	0	0	1	0
10	0	0	0	0	1
11	1	1	0	0	0
12	1	0	1	0	0
13	1	0	0	1	0
14	1	0	0	0	1
15	0	1	1	0	0
16	0	1	0	1	0
17	0	1	0	0	1
18	0	0	1	1	0
19	0	0	1	0	1
20	0	0	0	1	1
21	1	1	1	0	0
22	1	1	0	1	0
23	1	1	0	0	1
24	1	0	1	1	0
25	1	0	1	0	1
26	1	0	0	1	1
27	0	1	1	1	0
28	0	1	1	0	1
29	0	1	0	1	1
30	0	0	1	1	1

Note. The Q-matrix for the $J = 60$ conditions is doubled from this Q-matrix.

comparison of estimation variability, empirical *SE*s (i.e., standard deviations across replications) were computed.

The W, LR, and 2LR tests were computed for each dataset. In addition to assessing whether the 2LR test is a good approximation to the LR test, we also compared the performance of the 2LR and W tests in terms of Type I error and power. Type I error was computed as the proportion of times H_0 was rejected when the fitted reduced model is true; power was computed as the proportion of times that a wrong reduced model was rejected. The significance level was set at .05. With 200 replicates, the 95% confidence interval for the Type I error is given by $.05 \pm 1.96\sqrt{.05(1-.05)/200} = [.02, .08]$. A conservative performance (i.e., Type I error < .02) might be a good characteristic provided the power is not affected. For the purposes of this work, a statistical procedure was considered to be "good" if it had a Type I error within the [0, .08] interval and a relatively high power (> .80). The code used in this article was written in R. Some functions included in the GDINA (Ma & de la Torre, 2016) package were employed. The R code can be requested by contacting the corresponding author.

Results

Two-Step Estimation Procedure

Due to space limits, only the results of comparison in the worst ($N = 500$, $J = 30$, and IQ = LD) and best conditions ($N = 2{,}000$, $J = 60$, and IQ = HD) are presented in Tables 2 and 3. Items 1, 11, and 21, and 1, 21, and 41 are selected for $J = 30$ and $J = 60$, respectively. These are the same items that represent $K_j^* = 1$, 2, and 3 for both Q-matrices. In the case of the DINA model, we study the recovery of the probability of correct response in the two possible latent groups (i.e., φ_{j0} and $\varphi_{j0} + \varphi_{j12\ldots K_j^*}$). In the case of the *A*-CDM, we study the recovery of the baseline probability and the probability of correct response for examinees mastering only the first attribute (i.e., φ_{j0} and $\varphi_{j0} + \varphi_{j1}$).

In the worst condition, differences in terms of bias and empirical *SE* between the two algorithms were small, ranging from −.015 to .046, .010 being the mean and 0.012 the standard deviation. Not surprisingly, there was almost no difference between the two algorithms in the best condition – the largest absolute difference was 0.001. Considering both conditions, we can safely conclude that the differences of estimation accuracy and variability between the EM-MMLE and two-step algorithms were negligible. It should be noted that empirical SEs associated to the *A*-CDM estimates were usually larger compared to the DINA estimates. For example, this can be observed for the two-step estimates for item 21 in the worst condition. Empirical *SE*s for the *A*-CDM probabilities were .067 and .101. In the same condition, empirical *SE*s for the DINA probabilities were .038 and .045.

Two-Step Likelihood Ratio Test

Descriptive Analysis

All the item fit statistics were highly correlated. The Pearson correlation coefficients ranged from .97 to .99. Average computing time was recorded separately for each statistic. As an example, we found that in one of the most extreme conditions (i.e., $N = 2{,}000$, $J = 60$, and IQ = LD) the LR and 2LR tests took 475.03 and 1.61 seconds per replicate, respectively. In other words, the 2LR test was 295 times faster than the LR test.

Table 2. Selected item estimates for the DINA model

					Bias		Empirical standard error	
N	*J*	IQ	Item	Estimation algorithm	φ_{j0}	$\varphi_{j0}+\varphi_{j12\cdots\mathbf{K}^*_j}$	φ_{j0}	$\varphi_{j0}+\varphi_{j12\cdots\mathbf{K}^*_j}$
2000	60	HD	1	EM-MMLE	.000	.001	.012	.008
				Two-step	.000	.001	.012	.008
			21	EM-MMLE	.001	.000	.009	.009
				Two-step	.001	.000	.009	.009
			41	EM-MMLE	−.001	.000	.009	.010
				Two-step	−.001	.000	.009	.010
500	30	LD	1	EM-MMLE	−.010	−.001	.073	.030
				Two-step	.036	.001	.081	.034
			11	EM-MMLE	.003	.001	.043	.031
				Two-step	.029	.008	.052	.039
			21	EM-MMLE	.001	−.002	.030	.039
				Two-step	.018	.003	.038	.045

Notes. Generating values for the probabilities in the low discrimination (high discrimination) conditions were .30 (.10) and .70 (.90) for φ_{j0} and $\varphi_{j0}+\varphi_{j12\cdots K^*_j}$, respectively. *N* = sample size; *J* = test length; IQ = item quality; HD = high discrimination; LD = low discrimination.

Table 3. Selected item estimates for the A-CDM

					Bias		Empirical standard error	
N	*J*	IQ	Item	Estimation algorithm	φ_{j0}	$\varphi_{j0}+\varphi_{j1}$	φ_{j0}	$\varphi_{j0}+\varphi_{j1}$
2000	60	HD	1	EM-MMLE	−.001	.001	.012	.007
				2-step	−.001	.001	.012	.007
			21	EM-MMLE	−.001	.000	.016	.022
				2-step	−.001	−.001	.016	.021
			41	EM-MMLE	.000	−.001	.017	.025
				2-step	.000	−.002	.017	.025
500	30	LD	1	EM-MMLE	.003	−.004	.082	.040
				2-step	.033	−.005	.086	.040
			11	EM-MMLE	.001	.002	.079	.097
				2-step	.022	.022	.075	.098
			21	EM-MMLE	−.008	−.010	.060	.116
				2-step	.004	.003	.067	.101

Notes. Generating values for the probabilities in the low discrimination (high discrimination) conditions were .30 (.10) for φ_{j0} and .70, .50, and .43 (.90, .50, and .37) for $\varphi_{j0}+\varphi_{j1}$ for items 1, 11, 21, respectively. *N* = sample size; *J* = test length; IQ = item quality; HD = high discrimination; LD = low discrimination.

Type I Error

Type I error study results are presented in Table 4. The LR and 2LR tests were generally preferable to the W test. Type I error for the 2LR test was very similar to that obtained for the LR test. With the exception of low discriminating items, the 2LR and LR tests had an acceptable Type I error. LR and 2LR Type I error was close to the nominal level with low quality items when the sample size and test length were large (N = 2,000 and J = 60). 2LR Type I error was particularly good for DINA generated data. It was the only one lower than the upper limit of the confidence interval with medium quality items and a small sample size and test length (N = 500 and J = 30). The W test generally required a larger sample size. For example, W Type I error rate was inflated with medium quality items and small sample size (N = 500 and 1,000).

Power

Power study results are presented in Table 5. Power results should always be interpreted with some caution because power comparisons require equal Type I error. More liberal tests have a higher power because they tend to overestimate the significance. Power for all statistics was always higher than 0.80 and close to 1.00 in the high and medium discrimination conditions. In the case of the low quality items conditions, a large number of examinees (i.e., 1,000 or 2,000) or items (i.e., 60) were needed to reach acceptable values (i.e., > 0.80). 2LR power tended to be

Table 4. Type I error of the item fit statistics (LR, 2LR, and W) for the DINA and A-CDM models

Factors			DINA			*A*-CDM		
IQ	*J*	*N*	LR	2LR	W	LR	2LR	W
HD	30	500	.066	.022	.053	.058	.029	.079
		1,000	.062	.022	.090	.054	.029	.063
		2,000	.058	.022	.069	.048	.027	.054
	60	500	.061	.017	.055	.052	.016	.061
		1,000	.060	.017	.076	.051	.016	.055
		2,000	.051	.015	.062	.047	.015	.049
MD	30	500	.101	.075	.163	.145	.110	.233
		1,000	.068	.065	.109	.074	.083	.116
		2,000	.062	.060	.078	.053	.079	.067
	60	500	.070	.026	.105	.069	.033	.098
		1,000	.061	.026	.079	.059	.034	.070
		2,000	.050	.020	.060	.054	.031	.057
LD	30	500	.358	.443	.595	.374	.235	.581
		1,000	.223	.334	.430	.297	.290	.519
		2,000	.131	.278	.235	.224	.316	.371
	60	500	.199	.131	.323	.302	.133	.418
		1,000	.101	.096	.190	.156	.144	.262
		2,000	.071	.082	.102	.075	.118	.116

Notes. Shaded cells correspond to values in the [.00, .08] interval. IQ = item quality; *J* = test length; *N* = sample size; LR = likelihood ratio test; 2LR = two-step likelihood ratio test; W = Wald test; HD = high discrimination; MD = medium discrimination; LD = low discrimination.

Table 5. Power of the item fit statistics (LR, 2LR, and W) for the DINA and A-CDM models

Factors			Generating, true model: DINA Fitted, false model: *A*-CDM			Generating, true model: *A*-CDM Fitted, false model: DINA		
IQ	*J*	*N*	LR	2LR	W	LR	2LR	W
HD	30	500	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		1,000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	1.000
		2,000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	60	500	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		1,000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		2,000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
MD	30	500	1.000	1.000	1.000	.860	**.956**	.938
		1,000	**1.000**	1.000	1.000	**.994**	**.999**	.996
		2,000	**1.000**	**1.000**	1.**000**	**1.000**	**1.000**	**1.000**
	60	500	**1.000**	**1.000**	1.000	**.971**	**.979**	.980
		1,000	**1.000**	**1.000**	1.**000**	**1.000**	**1.000**	**1.000**
		2,000	**1.000**	**1.000**	1.**000**	**1.000**	**1.000**	**1.000**
LD	30	500	.595	.748	.759	.526	.776	.799
		1,000	.819	.952	.905	.589	.893	.837
		2,000	.979	.999	.987	.721	.959	.892
	60	500	.835	.916	.914	.533	.706	.749
		1,000	.984	.996	.991	.722	.906	.849
		2,000	**1.000**	1.000	1.000	**.963**	.995	.975

Notes. Shaded cells correspond to values in the [.80, 1.00] interval. Values shown in bold correspond to conditions where the actual Type I error was within the [.00, .08] interval. IQ = item quality; *J* = test length; *N* = sample size; LR = likelihood ratio test; 2LR = two-step likelihood ratio test; W = Wald test; HD = high discrimination; MD = medium discrimination. LD = Low discrimination.

higher that of the LR and W tests. For example, this was usually the case in the medium item quality conditions. It should be noted that in these conditions the 2LR Type I error was within the [0, .08] interval. In addition, it is important to note that, in the case of *A*-CDM generated data, 2LR power was much higher than that of the LR test in the low quality conditions, being .68 and .87 the marginal means associated to the LR and 2LR tests, respectively.

Discussion

Model-fit has received greater attention in the recent CDM literature (e.g., Chen et al., 2013; de la Torre & Lee, 2013; Hansen, Cai, Monroe, & Li, 2016; Liu et al., 2016; Sorrel et al., 2017). This an important area of research because proper application of a statistical model requires the assessment of model-data fit. Different reduced CDMs with different assumptions have been proposed in the literature. For example, the DINA model assumes a conjunctive process in that only individuals who master all required attributes are expected to correctly answer the item and the *A*-CDM assumes that the different attributes measured by the item contribute independently to the probability of correctly answer the item. A critical concern is in selecting the most appropriate model for each item from the available CDMs. To do so, several tests for model comparison have been employed, which include LR and the W tests. Although it has been found in the CDM context that the LR test is relatively more robust than the W test (Sorrel et al., 2017), the current implementation of the LR test is very time consuming, given that it requires to calibrate many different models and compare them to the general model. For this reason, the W test is generally preferred (e.g., de la Torre et al., 2015; Ravand, 2016) and is the one implemented in the software available (e.g., the CDM and GDINA packages in R; Ma & de la Torre, 2016; Robitzsch, Kiefer, George, & Uenlue, 2016).

In this work, we introduce an efficient approximation to the LR test, 2LR, based on a two-step estimation procedure under the G-DINA model framework originally introduced by de la Torre and Chen (2011). Results indicate that this two-step estimation procedure is comparable in terms of estimation accuracy and variability to the standard procedure based on EM-MMLE. Mean absolute differences and empirical standard errors produced by the two algorithms were very similar, even in the worst conditions. This shows that the estimates based on the two-step estimation procedure can be used to develop the approximation to the LR test.

The simulation study results allow us to draw several conclusions about the performance of the LR, 2LR, and W tests. First, the LR and 2LR tests were highly correlated. The performance of the 2LR test was very similar to that of the LR test. However, the computation of the 2LR test was remarkably faster. Secondly, the LR and 2LR tests were found to perform better than the W test. Thirdly, there was a large effect of the item quality. Type I error was close to the nominal level when the item quality was medium or high. In the poor discriminating item conditions, Type I error was inflated but in the case of the LR and 2LR tests this could be compensated by increasing the number of items or the sample size. Power decreased in the poor discriminating conditions. It is noteworthy that 2LR power was the least affected in these conditions and tended to be high. In sum, the 2LR test can be recommended for use in empirical research.

Following are some of the limitations of the current study and several avenues for future research. First, although not considered here, there is an additional reason to prefer the LR test over the W test – the LR test does not require the standard errors of the item parameter estimates, whereas the W test does. Future studies should explore the advantages of this feature. Second, all items were simulated to have the same discrimination power. This might not be feasible in practice. Finally, we focus on the DINA and *A*-CDM models and five attributes. Future studies might manipulate the number of attributes and try to extend this results to other models such as the *deterministic inputs, noisy "or" gate* (DINO) model (Templin & Henson, 2006), the linear logistic model (LLM; Maris, 1999), and the *reduced reparameterized unified model* (R-RUM; Hartz, 2002).

Acknowledgments

This research was supported by Grant PSI2013-44300-P (Ministerio de Economia y Competitividad and European Social Fund).

References

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange Multiplier tests: An expository note. *American Statistician, 36*, 153–157. doi: 10.2307/2683166

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*, 123–140. doi: 10.1111/j.1745-3984.2012.00185.x

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199. doi: 10.1007/s11336-011-9207-7

de la Torre, J., & Chen, J. (2011, April). *Estimating different reduced cognitive diagnosis models using a general framework*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*, 355–373. doi: 10.1111/jedm.12022

de la Torre, J., van der Ark, L. A., & Rossi, G. (2015). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*. Advance Online Publication. doi: 10.1177/0748175615569110

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333–352. doi: 10.1111/j.1745-3984.1989.tb00336.x

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology, 69*, 225–252. doi: 10.1111/bmsp.12074

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing, 11*, 144–177. doi: 10.1080/15305058.2010.534571

Liu, Y., Tian, W., & Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics, 41*, 3–26. doi: 10.3102/1076998615621293

Ma, W., & de la Torre, J. (2016). *GDINA: The generalized DINA model framework*. R package version 0.13.0. Retrieved from http://CRAN.R-project.org/package=GDINA

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187–212. doi: 10.1007/BF02294535

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment, 34*, 782–799. doi: 10.1177/0734282915623053

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2016). *CDM: Cognitive Diagnosis Modeling*. R package version 4. 8-0. Retrieved from http://CRAN.R-project.org/package=CDM

Rojas, G., de la Torre, J., & Olea, J. (2012, April). *Choosing between general and specific cognitive diagnosis models when the sample size is small*. Paper presented at the meeting of the National Council on Measurement in Education, Vancouver, Canada.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*, 219–262. doi: 10.1080/15366360802490866

Sorrel, M. A., Abad, F. J., Olea, J., Barrada, J. R., & de la Torre, J. (2017). Inferential item fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement*. Advance online publication. doi: 10.1177/0146621617707510

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods, 19*, 506–532. doi: 10.1177/1094428116630065

Tatsuoka, K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305. doi: 10.1037/1082-989X.11.3.287

Published online June 1, 2017

Miguel A. Sorrel
Department of Social Psychology and Methodology
Universidad Autónoma de Madrid
Ciudad Universitaria de Cantoblanco
Madrid 28049
Spain
miguel.sorrel@uam.es

Chapter 5

Model Comparison as a Way of Improving Cognitive Diagnosis Computerized Adaptive Testing

Abstract

Decisions on how to calibrate an item bank might have major implications in the subsequent performance of the adaptive algorithms. One of these decisions is model selection, which can be become problematic in the context of cognitive diagnosis computerized adaptive testing given the wide range of models available. This paper aims to determine whether model selection indices can be used to improve the performance of adaptive tests. Three factors were considered in a simulation study, i.e. calibration sample size, Q-matrix complexity, and item bank size. Results based on the true item parameters, and general and single reduced model parameter estimates were compared to those of the combination of appropriate reduced models within the generalized deterministic inputs, noisy, "and" gate model framework. The main implications of the current study for practical settings include an improvement in terms of classification accuracy and, consequently, testing time, and a more efficient use of the item pool.

Keywords: cognitive diagnosis models, computerized adaptive testing, model comparison, G-DINA, classification accuracy, item usage.

5.1 Introduction to the study

Adaptive testing methodologies, originally developed in the context of traditional item response theory (IRT), are being generalized to more complex scenarios, including cognitive diagnostic computerized adaptive testing (CD-CAT) (for a review, see Akbay and Kaplan 2017; Huebner 2010). CD-CAT is based on cognitive diagnosis models (CDMs), which are confirmatory latent trait models specifically developed

to detect mastery and nonmastery of set of fine-grained skills in a particular content domain. Some of the decisions that will affect the performance of the adaptive algorithms in this context involve the internal structure of the test specified in the Q-matrix and model selection. The identification of the Q-matrix is a laborious process where many professionals are typically involved (e.g., Li and Suen, 2013; Liu et al., 2013; Tjoe and de la Torre, 2014). For example, in the development of a Q-matrix for a proportional reasoning test, a diverse group composed of four mathematics researchers, three mathematics educations, five middle school mathematics teachers, and five graduate students in psychometrics and mathematics education was involved in Tjoe and de la Torre (2014). Examinees are also generally considered using think-aloud protocols to validate the theoretical framework (e.g., Li and Suen, 2013). The usual next step is to evaluate the initial Q-matrix using empirical Q-matrix validation procedures and evaluating the fit of the difference Q-matrix specifications (e.g., de la Torre and Chiu, 2016; Sorrel et al., 2016).

Arguably, the element that has received less attention is model selection. How to choose among the wide range of CDMs available is not an easy decision. Each CDM assumes a different cognitive process involved in responding to an item (e.g., conjunctive, disjunctive, or additive condensation rules). Besides, many CDMs have been created ranging in complexity. In this sense, recent developments have produced general CDMs such as the generalized deterministic inputs, noisy, "and" gate (G-DINA; de la Torre 2011) model, the general diagnostic model (GDM; von Davier 2005), and the log-linear CDM (LCDM; Henson et al. 2009). Reduced models are nested within these general models. Examples of reduced CDMs are the deterministic input, noisy "and" gate (DINA; Haertel 1989; Junker and Sijtsma 2001) model, the deterministic input, noisy "or" gate (DINO; Templin and Henson 2006), and the additive CDM (A-CDM; de la Torre 2011). Due to its relative novelty, CD-CAT empirical applications are still scarce. A trend may be noted, however, towards the use of the same CDM for all the items in the item bank. For example, Liu et al. (2013) applied a noncompensatory CDM, the DINA model, to a 352-item English language proficiency item bank. Reduced models have been widely used by researchers because of its simplicity of estimation and interpretation. However, these models make strong assumptions on the data and that's why their fit to the actual data should be evaluated.

A different alternative would be estimating a general model, allowing for all types of condensation rules within the same test. In this sense, Sorrel et al. (2018c) applied a generalization of the DINA model, the G-DINA model, to a 76-item proportional reasoning item bank. This alternative might be more consistent with results of real test data using CDMs indicating that no one model can be deemed appropriate for all test items (e.g., de la Torre and Lee, 2013; de la Torre et al., 2017; Ravand, 2016). Nevertheless, general models are more affected by the data conditions (e.g., need larger samples to be estimated accurately), and the risk of capitalization of chance is higher. Because of these shortcomings, researchers introduced item-level model comparison indices like the likelihood ratio (LR) for the purpose

of relative fit evaluation (Sorrel et al., 2017a). This allows for an intermediate situation between the two extremes (i.e., single reduced CDM vs. general model). The idea is to select the most appropriate CDM for each item. A further development on the LR test, the two-step LR test (2LR) demonstrated promise as a tool for assessing item relative fit in CDMs (Sorrel et al., 2017b). Importantly, the 2LR test is expected to perform very well under the usual item bank calibration conditions typically involving a large number of items (Sorrel et al., 2017a,b).

According to previous research, model selection might have an impact on classification accuracy (Ma et al., 2016; Rojas et al., 2012), and the generalization of the item parameter estimates (Olea et al., 2012). In this respect, Rojas et al. (2012) found that single reduced models, when appropriate, led to a better classification accuracy compared to general models. This was more notable in poor-quality conditions where it was more difficult to estimate the general model (e.g., small sample size and low item discrimination). In the context of IRT, Olea et al. (2012) explored the consequences of fitting a model under poor-quality item bank calibration conditions. They found that a parameters of the three-parameter logistic model were overestimated, causing an overestimation of the precision of the trait level estimates.

Therefore, all together, previous research indicates that test calibration conditions are highly related to the accuracy of the trait level estimates. This would be of major importance in the context of adapting testing where items are selected based on their parameter estimates. Considering all above, the present study investigates whether item-level model comparison indices can be useful to improve CD-CATs performance in terms of classification accuracy and item usage. The rest of the article is structured as follows. First, a detailed overview on CDM, item-level model comparison, and CD-CAT is provided. Second, the design of the simulation study is described, and the results under the different conditions are presented. Finally, in the discussion section, several implications and limitations of this study are discussed, and possible future directions are provided.

5.1.1 Cognitive diagnosis modeling

CDMs are confirmatory latent class models that are receiving increasing attention in the literature (for an overview of these models see, e.g., Rupp and Templin 2008; DiBello et al. 2007). Compared to traditional IRT where the underlying latent traits are continuous, in CDM the latent traits are discrete. Typically, these latent traits, commonly referred to as attributes, only have two levels: mastery and nonmastery. The goal of CDM is then to classify respondents as masters or nonmasters of a set of prespecified list of K attributes (e.g., skills, cognitive processes, disorders). This latent attribute vector or latent class can be denoted by α_l, for $l = 1, \ldots, 2^K$. These models emerged in the field of education (e.g., Tatsuoka 1990) and have been also applied in other settings such as clinical psychology and competency modeling (e.g., Templin and Henson, 2006; Sorrel et al., 2016). Multiple models have been proposed. Most of these

models are represented in general CDMs, such as the above-mentioned G-DINA model. This models partitions the latent classes into $2^{K_j^*}$ latent groups, where K_j^* is the number of attributes being measured by item j. The item response function (IRF) of the G-DINA model is then given by

$$P(X_j = 1|\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk}\alpha_{lk}\alpha_{lk'} \ldots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \tag{5.1}$$

where δ_0 is the intercept for item j, δ_{jk} is the main effect due to α_{lk}, $\delta_{jkk'}$ is the interaction effect due to α_{lk} and $\alpha_{lk'}$, and $\delta_{j12\ldots K_j^*}$ is the interaction effect due to $\alpha_1, \ldots, \alpha_{K_j^*}$. Thus, there are $2^{K_j^*}$ parameters to be estimated for item j.

Reduced CDMs are commonly encountered in the real data applications (e.g., Liu et al., 2013; Templin and Henson, 2006). They are formed by constraining some of the parameters of this general model. In this paper we consider three of these reduced models: DINA, DINO, and A-CDM. In the DINA model, all terms in Equation 5.1 except to the baseline probability and the highest interaction term are set to 0. The IRF for the DINA model is then given by

$$P(X_j = 1|\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \tag{5.2}$$

The DINA model only has two parameters per item: the guessing parameter represented by δ_0, and the slip parameter represented by $1 - \delta_{j0} + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}$. This is a noncompensatory model where the highest probability of success is only achieved when all the attributes required by the item have been mastered. On the contrary, the DINO model is a compensatory model. There are also only two parameters per item, namely δ_{j0} and δ_{j1}, with the important exception that δ_{j1} is constrained so that some lower-order terms will be cancelled by the corresponding high-order terms (de la Torre, 2011). The DINO model can be given by

$$P(X_j = 1|\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \delta_{j1} I(\boldsymbol{\alpha}_{lj}^* \neq \mathbf{0}) \tag{5.3}$$

where $I(\cdot)$ is an indicator variable. Respondents will have a high probability of success provided they master at least one required attribute.

Finally, the A-CDM is an additive model where all the interaction terms are dropped, and thus each mastered attribute contributes to the probability of success indepently. The IRF of the A-CDM is given by

$$P(X_j = 1|\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} \tag{5.4}$$

Given that all these reduced models are nested within the more general G-DINA model, item-level model comparison statistics can be computed to compare the relative fit of the different models. The approach that is considered in this paper is detailed in the following section.

5.1.2 Item-level model comparison

Among all the competing models, the Occam's razor principle dictates that the simplest should be chosen. One of the reasons for doing that is to avoid the capitalization on chance problem. Different studies pointed out that particularly when the sample size is small and the item quality is poor, an appropriate reduced CDM will lead to a higher accuracy (Ma et al., 2016; Rojas et al., 2012). The advantage of using a reduced CDM will be greater the more complex the item structure. In the extreme case of minimum complexity, it is irrelevant which CDM is applied because all of them are equivalent. Two fictitious items are depicted in Table 1 for illustration purposes. Item A is a one-attribute item and item B is a three-attribute item. As can be seen from the table, one-attribute items do not follow any particular CDM in the sense that all models only have two parameters, namely the baseline probability and the additive effect of mastering the attribute. In contrast, more complex items such as item B will lead to different IRF specifications according to the different CDMs. In this sense, the DINA and DINO models will always have two parameters per item regardless of the item complexity, but the number of item parameters will linearly and exponentially grow for the A-CDM and G-DINA models, respectively. As such, 4 and 8 parameters are estimated in the case of item B. Sample size requirements for estimation of item parameters in complex structures will be stricter. If the number of attributes being measured by the item is high (e.g., 3-4 attributes), then these parameters will not be accurately estimated unless the sample size is high.

Several statistics have been proposed for the purpose of assessing relative fit in the context of CDM (for a comparison of these tests see, e.g., Sorrel et al., 2017a,b). The nice feature of the item-level model comparison tests is that they allow selecting the most appropriate reduced model for each item, considerably decreasing the number of parameters to be estimated in the case of complex items. One of these item-level model comparison tests is the LR test. The traditional implementation of the LR test requires both the general and the reduced CDM to be estimated. Sorrel et al. (2017b) proposed an approximation with the advantage of only requiring the more general model to be estimated, referred to as two-step LR test (2LR). Item-level maximum likelihoods for the competing models are estimated using the following formula:

$$L(\boldsymbol{\delta}_j|\boldsymbol{R}_j,\boldsymbol{I}_j) = \sum_{k=1}^{2^{K_j^*}} P^{(m)}(\boldsymbol{\alpha}_{lj}^*)^{R_{\alpha_{lj}}}[1-P^{(m)}(\boldsymbol{\alpha}_{lj}^*)]^{I_{\alpha_{lj}}-R_{\alpha_{lj}}} \tag{5.5}$$

where $P^{(m)}(\boldsymbol{\alpha}_{lj}^*)$ represents the probability of correctly answering the item j for respondents in the latent group l based on the item parameters of the model of interest, and $I_{\alpha_{lj}}$ and $R_{\alpha_{lj}}$ represents the

Table 5.1: Item parameters for items measuring one and three attributes

Item A with q-vector = {10000}								
Models	$P(0)$	$P(1)$						
G-DINA	δ_0	$\delta_0+\delta_1$						
DINA	δ_0	$\delta_0+\delta_1$						
DINO	δ_0	$\delta_0+\delta_1$						
A-CDM	δ_0	$\delta_0+\delta_1$						
Item B with q-vector = {11100}								
Models	$P(000)$	$P(100)$	$P(010)$	$P(001)$	$P(110)$	$P(101)$	$P(011)$	$P(111)$
G-DINA	δ_0	$\delta_0+\delta_1$	$\delta_0+\delta_2$	$\delta_0+\delta_3$	$\delta_0+\delta_1+\delta_2+\delta_{12}$	$\delta_0+\delta_1+\delta_3+\delta_{13}$	$\delta_0+\delta_2+\delta_3+\delta_{23}$	$\delta_0+\delta_1+\delta_2+\delta_3+\delta_{12}+\delta_{13}+\delta_{23}+\delta_{123}$
DINA	δ_0	δ_0	δ_0	δ_0	δ_0	δ_0	δ_0	$\delta_0+\delta_1$
DINO	δ_0	$\delta_0+\delta_1$	$\delta_0+\delta_1$	$\delta_0+\delta_1$	$\delta_0+\delta_1$	$\delta_0+\delta_1$	$\delta_0+\delta_1$	$\delta_0+\delta_1$
A-CDM	δ_0	$\delta_0+\delta_1$	$\delta_0+\delta_2$	$\delta_0+\delta_3$	$\delta_0+\delta_1+\delta_2$	$\delta_0+\delta_1+\delta_3$	$\delta_0+\delta_2+\delta_3$	$\delta_0+\delta_1+\delta_2+\delta_3$

number of respondents in latent group l and the number of respondents in latent group l who correctly answered the item based on the attribute joint distribution estimated for the more general model (i.e., G-DINA). This statistic performed well under different scenarios of sample size, test length, generating model, and item quality. In the CD-CAT conditions, where the item bank length is typically large, item-level relative fit statistics are expected to perform very well taking into account the effect of the test length on its performance (Sorrel et al., 2017a,b).

5.1.3 Cognitive diagnosis computerized adaptive testing

The area of application of CDM to adaptive testing is referred to as CD-CAT. This is a new area of application that has been aided from the developments in traditional CAT, typically based on IRT. Unfortunately, because latent variables in CDMs are discrete, item selection methods based on the Fisher information cannot be applied in CD-CAT. However, several item selection indices have been proposed for CD-CAT, including the generalized deterministic inputs, noisy "and" gate model discrimination index (GDI; Kaplan et al. 2015). This index yielded shorter test administration times compared to the other item selection methods (e.g., modified posterior weighted Kullback-Leibler). The next item to be selected by the adaptive algorithm is the one with the highest GDI:

$$s = \text{argmax}_{j \in B_q} GDI = \sum_{k=1}^{2^{K_j^*}} \pi(\boldsymbol{\alpha}_{lj}^*)[P(\boldsymbol{\alpha}_{lj}^*) - \bar{P}_j]^2 \tag{5.6}$$

where $\boldsymbol{\alpha}_{lj}^*$ defines a reduced attribute pattern, $\pi(\boldsymbol{\alpha}_{lj}^*)$ the probability of $\boldsymbol{\alpha}_{lj}^*$, $P(\boldsymbol{\alpha}_{lj}^*)$ the conditional probability of success on item j given by the reduced latent pattern $\boldsymbol{\alpha}_{lj}^*$, and $\bar{P}_j$ the average success probability is computed as $\bar{P}_j = \sum_{k=1}^{2^{K_j^*}} \pi(\boldsymbol{\alpha}_{lj}^*)P(\boldsymbol{\alpha}_{lj}^*)$.

5.1.4 Goal of the present study

The present study aims to explore the impact of item bank calibration on the CD-CAT performance. Specifically, it is assessed to what extent a better performance can be obtained when appropriate reduced CDMs are estimated for each item using an item-level model comparison index, namely the 2LR test. Hypothetically, the 2LR test will show a very good performance under the usual item bank calibration conditions (e.g., large pool of items). Thus, it is expected that this index will be useful in improving CD-CAT performance. Compared to a situation where a general model is estimated for all the items, a combination of models derived by the 2LR test will require estimating fewer parameters. Thus, these parameters will be estimated more accurately, having an impact on the classification accuracy. In addition, item usage under the different item bank calibration conditions will be explored. These research questions are addressed by using Monte Carlo Methods. Only low item discrimination conditions are considered

because a larger improvement can be expected in these situations given that item parameters are estimated less accurately in these situations (Ma et al., 2016; Rojas et al., 2012).

5.2 Method

A simulation study was conducted to evaluate the classification accuracy and item usage obtained with each of the model calibrations described: G-DINA, 2LR-derived combination of models, DINA, DINO, and A-CDM. For comparison purposes, true item parameters were also considered, which allows obtaining an estimation of the upper limit for the classification accuracy. Factors and levels were selected based on a literature review of current CDM and CD-CAT empirical applications (e.g., Liu et al., 2013; Sorrel et al., 2016, 2018c). Three data factors were varied, namely the calibration sample size (N = 250, 500, and 2,000 respondents), the item bank length (J = 155 and 310 items), and the Q-matrix complexity ($Q - str$ = simple and complex Q-matrix structure). More specifically, Q-matrix complexity was understood as the number of attributes being measuring by each of the items (see Table 1 where this concept was discussed). Two levels were considered for this factor. In the simple Q-matrix condition, 35 one-, 60 two-, and 60 three-attribute items were generated. On the contrary, in the complex Q-matrix condition, 60 two-, 60 three-, and 30 four attribute items were generated, and 5 additional one-attribute items conditions were also included to ensure completeness of the Q-matrix, which is a necessary condition for the identifiability of the population proportion parameters (Gu and Xu, 2017; Xu and Zhang, 2016). In the J = 310 item conditions, these numbers are doubled.

The 2LR test is an inferential test and thus a significance level needs to be selected. We report the results for $\alpha = .05$. In addition, we implemented the Bonferroni (BF) correction considering the fact a large number of tests was being considered. For example, in the 155 items and simple Q-matrix condition, there were 120 items measuring more than one attribute. Given that we considered three possible reduced models (i.e., DINA, DINO, and A-CDM), 360 tests were conducted. The Bonferroni correction sets the significance cut-off at α/t, being t the number of tests. Although the BF correction tends to be a bit too conservative, this is not a big issue here because we retained the reduced model with the highest p-value associated among the models with a p-value greater than .05.

Ten item banks were constructed for each simulated condition. Item parameters were generated randomly from the following distributions: $P(\mathbf{0}) \sim U(0.20, 0.40)$ and $P(\mathbf{1}) \sim U(0.60, 0.80)$. For the A-CDM model, the main effect of each attribute was set to be $P(\mathbf{0}) + (P(\mathbf{1} - P(\mathbf{0}))/K_j^*$, where K_j^* denotes the number of attributes being measured by the item. This condition has been referred to as low item quality in previous research (e.g., Ma et al., 2016; Sorrel et al., 2017a,b). Finally, some other factors were also fixed. The number of attributes was fixed to 5, which is a reasonable value considering current CDM empirical applications (e.g., Li et al., 2016; Ravand, 2016; Sorrel et al., 2016) and simulation studies

(e.g., Cheng, 2009, 2010; Kaplan et al., 2015). The generating, true model used in the data generation process was always a combination of the same number of DINA, DINO, and A-CDM items. It should be remarked here all the different CDMs are equivalent when the number of attributes being measured by the item is one. We included 40 and 50 items for each reduced model in the simple and complex Q-matrix conditions, respectively. In the J = 310 item conditions, these numbers are doubled.

For each condition and item bank, we generate a validation sample consisting of 5,000 response patterns generated uniformly from the space of possible $2^5 = 32$ latent classes. The following are details on CD-CAT simulation. The first item was randomly chosen from the medium discriminating items. The item selection rule was the generalized discrimination index (GDI; Kaplan et al. 2015). The descriptive statistics for GDI in one of the generated conditions was: $mean = 0.026$, $SD = 0.017$, $min = 0.005$, and $max = 0.087$. The scoring method was the maximum a posteriori (MAP) method. The MAP estimate of examinee i is given by (Huebner and Wang, 2011):

$$\hat{\boldsymbol{\alpha}}_{MAP} = \text{argmax}_l \, P(\boldsymbol{\alpha}_l | \boldsymbol{X}_i) \tag{5.7}$$

where the posterior probability $P(\boldsymbol{\alpha}_l | \boldsymbol{X}_i)$ is computed using Bayes' theorem.

Two dependent variables were used for the comparison between the different CD-CAT applications: pattern recovery (i.e., proportion correctly classified vectors), and item usage for the different item types. Pattern recovery was computed as:

$$PCV = \frac{\sum_{i=1}^{N} I[\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i]}{N} \tag{5.8}$$

where $I[\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i]$ evaluates whether the estimated attribute vectors matches the generated attributes. Item usage was computed from the item exposure rates. A total of 150,000 items was administered for each test bank replication (i.e., 30 items for each of the 5,000 examinees). We computed the number of items administered within each item type category (i.e., q-vector complexity: one-, two, three-, and four-attribute items, and generating model: DINA, DINO, and A-CDM), and divided theses sums by 150,000 then obtaining an indicator of item usage that is relative to the total number of items administered (i.e., different mutually exclusive categories sum up to 1). Model estimation was conducted using the **GDINA** R package (Ma and de la Torre, 2017). Another code that can be requested by contacting the corresponding author was written in R for the 2LR test and CD-CAT analyses.

5.3 Results

5.3.1 Calibration sample results: Model selection

Table 2 [1] includes the results for the average 2LR performance across all the test bank replications. As can be seen from the table, even in the small calibration sample size conditions, the true reduced model was selected at least in 68% of the items. Small sample size conditions affected the power of the statistic, thus increasing the number of times that an incorrect reduced CDM was retained. As the calibration sample size increased, the selection rates improved. In the N = 2000 condition, the true reduced model was selected at least in 90% of the items. The large number of comparisons caused that the Type I error rate increased (i.e., the probability of rejecting that the generating reduced model does not fit the data as well as the more general model). Accordingly, results for the BF correction were always better.

The good performance of the 2LR test allowed dramatically reducing the number of parameters to be estimated. This is illustrated in Table 3, where the average number of parameters estimated under different conditions is included. For example, in the J = 310 and simple Q-matrix condition, the GDINA model estimated 1580 parameters, whereas the combination of models selected by the 2LR test estimated an average of 728.20 to 739.60 for different levels of calibration sample size across all the test bank replications. Thus, using the 2LR test led to a reduction of approximately 50% in the number of parameters to be estimated. This might have a notable impact on the accuracy of those item parameter estimates, affecting the performance of the adaptive algorithms based on those estimates, as we explore in the next section. Though generally rare, there were sometimes an incorrect reduced model was selected by 2LR. This explains why sometimes the average number of parameters is higher in large sample size conditions, compared to small sample size conditions.

5.3.2 Validation sample results: Pattern recovery

Pattern recovery results are shown in Figures 5.1 and 5.2 for the 155 and 310 item bank length conditions, respectively. For comparison purposes, the upper-limit of the pattern recovery (i.e., the one that is obtained when the true item parameters are used) is represented in black. The goal of this section is to explore whether the different model calibrations are close to that upper-limit. Conditional results on different CD-CAT length conditions are depicted: starting from only one item up to 30 items. However, most of the results will be described assuming that the CD-CAT length was fixed to 30 items, a reasonable test length that provides sufficient classification accuracy considering prior research (Kaplan et al., 2015). In the following we describe the most notable findings.

[1] Results in terms of selection rate, pattern recovery, and item usage based on a different item-level model comparison index, the Wald test, were essentially the same. However, the 2LR test was found to be faster than the Wald test in terms of implementation time. For example, in the N = 2000, J = 155, and simple Q-matrix condition 2LR and W analyses took 8 seconds and 17 minutes, respectively.

Table 5.2: Model selection rates for the 2LR test.

			Calibration Sample Size								
			N = 250			N = 500			N = 2,000		
Item Bank Length	Multiple comparison correction	Q-Matrix Structure	Correct	G-DINA	Incorrect	Correct	G-DINA	Incorrect	Correct	G-DINA	Incorrect
J = 155	None ($\alpha = .05$)	Simple	**0.82**	0.06	0.12	**0.87**	0.07	0.06	**0.93**	0.07	0.00
		Complex	0.68	0.12	0.20	0.79	0.13	0.08	**0.90**	0.10	0.00
	BF Correction	Simple	**0.87**	0.00	0.13	**0.94**	0.00	0.06	**0.99**	0.00	0.01
		Complex	0.78	0.00	0.22	**0.91**	0.00	0.09	**0.99**	0.00	0.00
J = 310	None ($\alpha = .05$)	Simple	**0.84**	0.04	0.11	**0.92**	0.04	0.04	**0.94**	0.06	0.00
		Complex	0.78	0.04	0.18	**0.89**	0.04	0.07	**0.95**	0.05	0.055
	BF Correction	Simple	**0.88**	0.00	0.12	**0.96**	0.00	0.04	**1.00**	0.00	0.00
		Complex	**0.81**	0.00	0.19	**0.93**	0.00	0.07	**1.00**	0.00	0.01

Note. There are 35 and 5 one-attribute items for which the 2LR test is not performed in the simple and complex Q-matrix structure conditions, respectively. There are 120 and 150 tests performed under each condition, respectively. In the J = 310 conditions these numbers are doubled. Cells with values higher than 0.80 are shown in bold. Correct: the true, generating reduced CDM is retained. G-DINA: the G-DINA model is retained. Incorrect: a false, nongenerating reduced CDM is retained.

Table 5.3: Number of parameters estimated by the G-DINA model and average number of parameters estimated by the 2LR-derived combination of models.

			Calibration Sample Size		
Item Bank Length	Q-Matrix Structure	Model	N = 250	N = 500	N = 2,000
J = 155	Simple	GDINA		790	
		2LR - BF (average)	369.70	367.00	369.90
	Complex	G-DINA		1210	
		2LR - BF (average)	407.00	400.40	400.60
J = 310	Simple	G-DINA		1580	
		2LR - BF (average)	728.20	733.90	739.60
	Complex	G-DINA		2420	
		2LR - BF (average)	804.50	795.30	799.70

General vs. reduced CDMs. The true underlying model for the item bank was a combination of DINA, DINO, and A-CDM items. Thus, as expected, estimating the same reduced model (i.e., DINA, DINO, or A-CDM) for all the items in the item banks resulted in a poorer performance of the CD-CAT compared to that of the CD-CAT based on the G-DINA model that subsumes all of them. Among the reduced models, CD-CATs based on the DINA and DINO models performed similarly, and CD-CATs based on the A-CDM performed considerably worse in all conditions. As indicated in the previous section, the 2LR test generally flagged the more appropriate model for each item. Consequently, CD-CAT based on that combination of models usually had a very good overall performance. Indeed, the performance of this combination of models was always equal or better compared to that of the G-DINA model. For a 30-item CD-CAT, the average improvement in pattern recovery that was obtained when the 2LR test along with the BF correction across all conditions and replications was 4.76%, and ranged from 0.16 to 21.44%.

Multiple comparison correction. Including the BF correction always led to a better performance of the CD-CAT. This was related to the results described in the model selection section. Differences were more notable when the calibration sample the item bank length were small, and Q-matrix was complex. Differences were negligible, for example, when the calibration sample size and the item bank length were large.

Calibration sample size. Best results were always obtained when the true item parameters were employed. This was due to the sampling estimating error when estimating the item parameters under any condition. As expected, the sampling estimating error was smaller when the sample size was large (i.e., N = 2,000), and then the results for pattern recovery for the more general model (i.e., G-DINA; red line) were close to the upper limit. The same can be said for the combination of models selected by the 2LR test results given that this statistic performed very well under this condition. In contrast, the G-DINA model was not accurately estimated under small sample conditions (i.e., N = 500, 250), and that's why the CD-CAT based on the G-DINA model parameters performed poorly when the calibration sample size became smaller. As indicated in Table 2, the 2LR test did a good job in selecting the true generating reduced CDM, CD-CAT based on the 2LR test results then performed generally close to the upper limit given that reduced CDMs were easier to estimate under small calibration sample size conditions.

Q-matrix complexity. It was always harder to recover the attribute vector when the Q-matrix was complex. This decrement in accuracy was more pronounced for the G-DINA model as the number of model parameters to be estimated was higher. For example, in the complex Q-matrix for an item measuring four attributes $2^4 = 16$ parameters were estimated under the G-DINA model. If the DINA model fitted that particular item according to the 2LR test, only two parameters were estimated. When the sample size was large enough, no difference between the performance of the G-DINA and 2LR test selection of models was found. However, performance of the CD-CAT based on the G-DINA became worse as

the sample size decreased. In this line, even when the calibration sample size was 250, the CD-CAT based on the 2LR test selection was still relatively close to the upper limit. In the $J = 310$ condition, for a 30-item CD-CAT the average pattern recovery was 0.84, 0,77, and 0.58 for CD-CATs based on the true item parameters, 2LR test along with the BF correction selection and G-DINA estimates, respectively.

Item bank length. Increasing the item bank length always led to a better performance of the CD-CAT. Item parameters were generated using a uniform distribution, then increasing the number of items in the item bank made the probability of having more high quality items higher, and improving item bank calibration conditions and 2LR test selection rates. A horizontal reference line at pattern recovery = 0.70 is included in all figures. As can be observed, results based on the 2LR test selection of models always achieved that limit, except in the more problematic condition (i.e., small calibration sample size and item bank length, complex Q-matrix). In the $J = 310$ items condition, a 20-item CD-CAT was generally enough to achieve that limit.

5.3.3 Validation sample results: Item usage

Average item usage results across the ten item banks are shown in Table 4. Only the large item bank length condition is considered to prevent the different item types in the item bank from being exhausted by the selection algorithm. In addition, considering the space limits, only the most and least ideal data conditions are presented (i.e., simple vs. complex Q-matrix, large vs. small calibration sample size). The most notable results were the following:

- Simpler items were typically preferred when using GDI. It should be noted that in the complex Q-matrix condition there were only ten one-attribute items. Probably, highly discriminating one-attribute items were exhausted. Given that, it can be expected that the use of one-attribute items would be greater if a wider range of one-attribute items were available.

- The patterns of item usage for both $p = 0.05$ and BF implementations of the 2LR test were quite similar, and were the ones closest to the pattern corresponding to the true estimates. These patterns were also similar to that of the G-DINA model, although there some discrepancies in the complex Q-matrix condition. The G-DINA model tended to use more four-attribute items.

- When the data were calibrated using a single reduced CDM, one-attribute items were generally preferred. It should be noted that all the CDMs are equivalent when the number of attributes being measured by the item is one, all having only two parameters. In addition, when the data were calibrated using a single reduced CDM, items following a different model were seldom used, and items following that specific reduced models were mostly used. The former was more

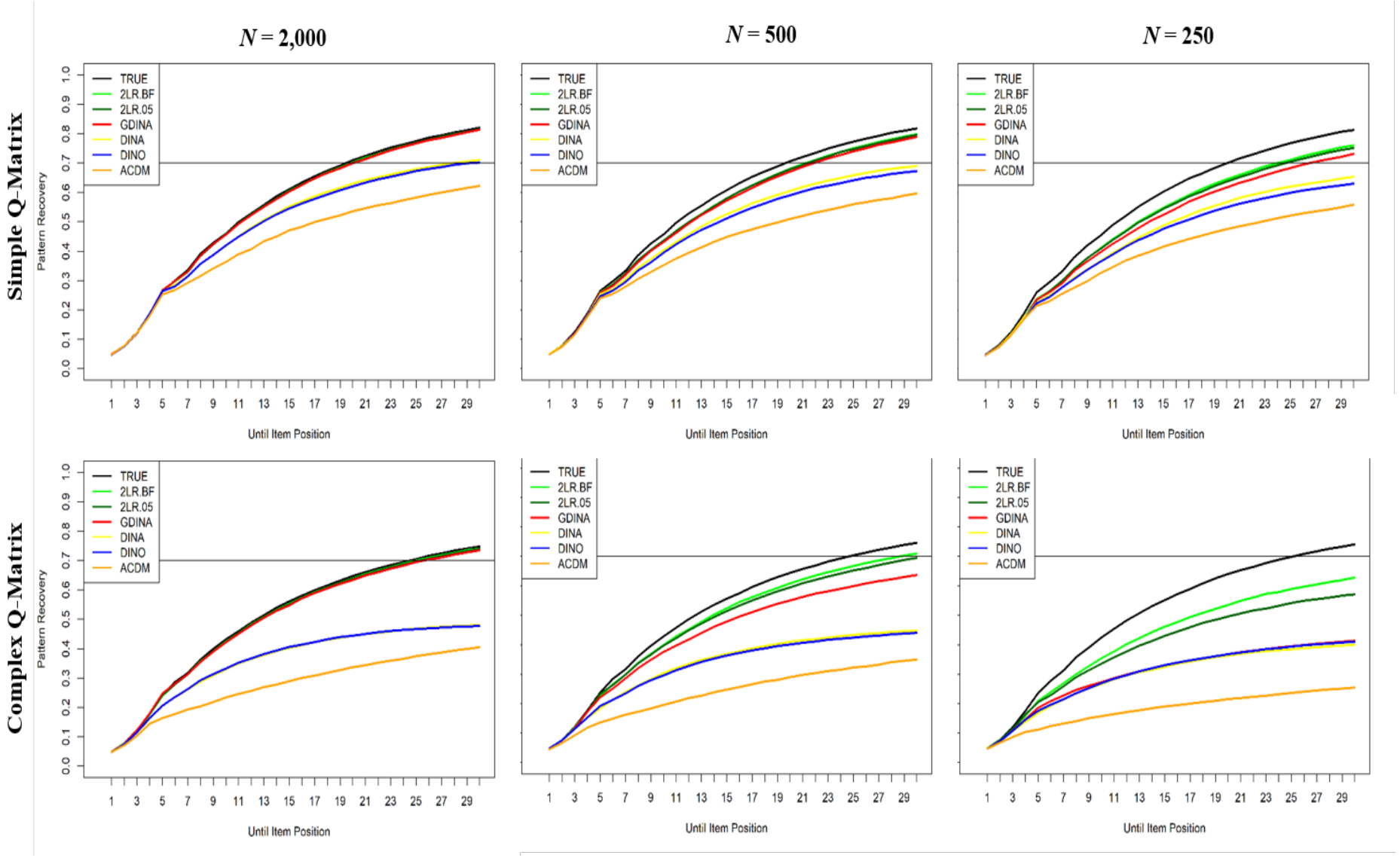

Figure 5.1: Pattern recovery according to fitted model and item position. Item bank length is 155 items. A horizontal line is including at 0.70 for interpretation purposes.

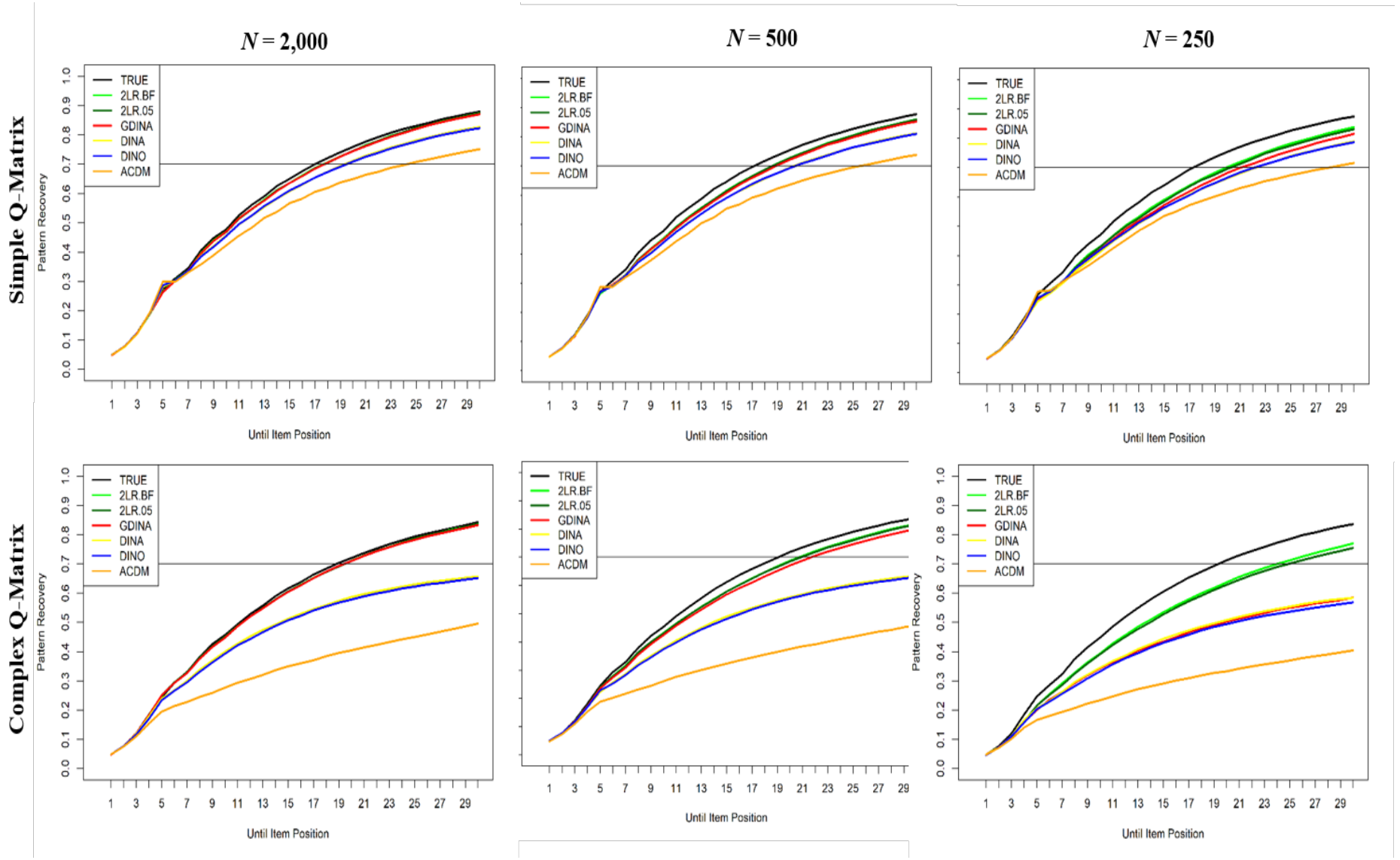

Figure 5.2: Pattern recovery according to fitted model and item position. Item bank length is 310 items. A horizontal line is including at 0.70 for interpretation purposes.

pronounced when Q-matrix was simple, whereas the later was more pronounced when the Q-matrix was complex.

- Items following the A-CDM model were seldom administered. This was most noticeable in the simple Q-matrix conditions, where, even when the item bank was calibrated using the A-CDM model, A-CDM items were almost not administered. In this situation, the adaptive procedure generally administered one-attribute items.

All above has to do with the fact that item parameters were not properly estimated when the calibrated reduced model was different from the true generating model. This is illustrated in Table 5 where the estimated parameters for two items in the N = 2000 and simple Q-matrix condition are presented. Item 3 was a one-attribute item, and thus all the CDMs provided essentially the same item parameters. This might explain why one-attribute items were usually used under any condition. On the contrary, Item 20 was a two-attribute item following the DINA model. As can be seen from the table, the estimated GDI for DINO and A-CDM was quite low, whereas the G-DINA model, the combination of models derived from the 2LR test, and the DINA model provided similar results, and were close to the GDI that was specified in the data generation.

5.4 Discussion

In current empirical studies using CD-CAT, a single reduced CDM is applied to all items in the item bank (e.g., Liu et al. 2013). Generally, this might not be a suitable approach given that reduced models make strong assumptions about the data, so they might not be appropriate for all items. Accounting for the heterogeneity of CDMs even within the same test found in some empirical studies (de la Torre and Lee, 2013; de la Torre et al., 2017; Ravand, 2016), the use of general CDMs emerged as a good alternative (Sorrel et al., 2018c). This alternative has the limitation that the estimation of general CDMs is much more challenging. General CDMs needs a larger sample size to be estimated accurately, and this requirements will be stricter and stricter as the number of parameters increase. Considering this, the present study explores whether the classification accuracy and the usage of the item bank be improved by using comparison indices to select the most appropriate model for each item. The results seem to indicate that implementing item-level model comparison indices such as 2LR test (Sorrel et al., 2017b) improved the accuracy of the CD-CAT under all the simulated conditions. Accordingly, the same accuracy can be obtained with fewer items administered. For example, in the most challenging simulated condition (i.e., small sample size, complex Q-matrix structure), a CD-CAT based on the 2LR test combination of models would need 25 items to achieve an accuracy of .70, whereas a CD-CAT based on the G-DINA achieved an accuracy of .60 by the time that the CD-CAT stopped after the administration of 30 items thus

Table 5.4: Average item usage results for the 310-item banks in the most and least ideal conditions.

Most Ideal Condition: $Q - str$ = Simple Structure & N = 2000						
Fitted Model	one-attribute	two-attribute	three-attribute	DINA-items	DINO-items	A-CDM-items
Number of Items	(#70)	(#120)	(#120)	(#80)	(#80)	(#80)
TRUE	**0.37**	**0.37**	0.26	**0.31**	**0.32**	0.00
2LR_BF	**0.36**	**0.37**	0.27	**0.32**	**0.32**	0.00
2LR_05	**0.35**	**0.37**	0.27	**0.32**	**0.33**	0.00
G-DINA	**0.35**	**0.37**	0.28	**0.32**	**0.33**	0.00
DINA	**0.59**	0.25	0.16	**0.39**	0.00	0.02
DINO	**0.59**	0.26	0.16	0.00	**0.40**	0.02
A-CDM	**0.93**	0.07	0.00	0.01	0.01	**0.06**

Least Ideal Condition: $Q - str$ = Complex Structure & N = 250							
Fitted Model	one-attribute	two-attribute	three-attribute	four-attribute	DINA-items	DINO-items	A-CDM-items
Number of Items	(#10)	(#120)	(#120)	(#60)	(#100)	(#100)	(#100)
TRUE	0.10	**0.54**	0.29	0.07	**0.45**	**0.44**	0.00
2LR_BF	0.10	**0.56**	0.28	0.07	**0.44**	**0.44**	0.03
2LR_05	0.09	**0.55**	0.29	0.07	**0.44**	**0.44**	0.03
G-DINA	0.09	**0.49**	0.28	0.15	**0.42**	**0.41**	0.09
DINA	0.21	**0.55**	0.19	0.05	**0.63**	0.02	0.15
DINO	0.20	**0.56**	0.19	0.05	0.02	**0.63**	0.15
A-CDM	0.24	**0.63**	0.12	0.02	0.20	0.18	**0.38**

$Note$. Maximum values within each item type category are shown in bold (± 0.02 differences are not considered).

requiring a much larger number of item. This time saving might be of major importance, for example, in classroom settings, because it would allow teachers designing classroom specific activities to optimize student learning (Chang, 2015; Shute et al., 2016). On the whole, the efficiency of adaptive testing can make assessment less intrusive, thus more practicable, in many different contexts where this is a concern (e.g., educational, medical).

Table 5.5: True and estimated item parameters for two different item types

One-Attribute Item: Item 157 with q-vector = {01000}					
Models	GDI	$P(0)$	$P(1)$		
TRUE	0.066	0.243	0.757		
2LR-BF	0.061	0.251	0.747		
2LR	0.061	0.251	0.747		
G-DINA	0.061	0.251	0.747		
DINA	0.059	0.250	0.737		
DINO	0.061	0.255	0.748		
A-CDM	0.063	0.249	0.749		
DINA Two-Attribute Item: Item 18 with q-vector = {00110}					
Models	GDI	$P(000)$	$P(100)$	$P(010)$	$P(001)$
TRUE	0.067	0.200	0.200	0.200	0.800
2LR-BF	0.068	0.195	0.195	0.195	0.798
2LR	0.068	0.195	0.195	0.195	0.799
G-DINA	0.068	0.190	0.172	0.222	0.799
DINA	0.068	0.193	0.193	0.193	0.795
DINO	**0.009**	0.185	**0.399**	**0.399**	**0.399**
A-CDM	**0.032**	**0.114**	**0.332**	**0.402**	**0.620**

Note. Differences with respect to the true values greater than 0.03 are shown in bold.

Regarding the manipulated factors, we found that the accuracy improvement can be expected to be larger when the Q-matrix structure is complex (i.e., large proportion of items measuring more than one attribute) and the calibration sample size is small. Otherwise, if the same reduced CDM (e.g., DINA) is applied to all items in a situation in which items follow several different CDMs, the resulting accuracy will be generally much lower. This might be ameliorated in a certain way if the sample size is large and the Q-matrix has a simple structure, as in Liu et al. (2013), but still a CD-CAT based on a general model of a combination of models would provide better accuracy results. Furthermore, even if a similar accuracy is obtained with the application of a single reduced model, there will be a poor use the item bank. Specifically, items following a different reduced CDM won't be selected by the adaptive procedure. This is due to a severe underestimation of the model discrimination when an incorrect reduced CDM is specified for an item. Results of the current study indicate that this inefficient use of the item bank can be tackled through the use of model selection indices. On the other hand, procedures based on a general model (e.g., G-DINA) will lead to optimal results provided the general model is accurately estimated. This will generally be the case when the sample size is large and the number of parameters to be estimated

is small (e.g., Sorrel et al. 2018c). Otherwise the classification accuracy can be much lower, compared to implementation of the model comparison indices.

In any case, the main finding of this study is that we can improve classification accuracy and make a better use of the item bank by using item-level model fit indices to select the most appropriate CDMs for each item. Importantly, it will not have any negative impact. This study considers DINA, DINO, and A-CDM models, but different constrained versions of the G-DINA model can be easily included in the set of models (e.g., R-RUM; Hartz 2002; LLM; Maris 1999). Given the large number of comparisons, researches and practitioners are encourage to use a procedure to control de Type I error rate such as the Bonferroni correction, as it was done here. It is worth noting that these methodologies are indeed very easy to implement. In this sense, it only took a few seconds for an item bank composed of 155 items to conduct the model selection analysis. The code for the 2LR test can be requested from the corresponding author. In addition, **CDM** and **GDINA** R packages (Ma and de la Torre, 2017; Robitzsch et al., 2017) include functions to compute similar statistics.

Findings from this study can serve future research in several ways. First, we found that simpler items were typically preferred by GDI. This was more noticeable in the complex Q-matrix condition. One of the possible reasons is that all models are equivalent when the item measures only one attribute, whereas the models are more and more different as the item complexity increases. If the appropriate reduced model is not correctly specified, then the item discrimination would be severely underestimated. This result is in line with previous research using GDI (Kaplan et al., 2015; Yigit et al., 2018). On another note, items following A-CDM were not generally administered, which might lead to explore more deeply the formulation of GDI as item selection rule. The GDI measures the weighted variance of the probabilities of success of an item. For illustrative purposes, Figure 5.3 includes the success probabilities of each latent group for a fictitious three-attribute item. Considering that in this example the Q-matrix is composed only of three attributes, the DINA, DINO, and A-CDM will have 2, 2, and 3 possible latent groups, respectively. If we compute the GDI (detailed in Equation 5.6) using a uniform prior for the attribute joint distribution, the resulting GDI will be 0.03 for the DINA and DINO models, and 0.02 for the A-CDM model. These item parameter estimates would considered as low discrimination in terms of the guessing and slipping parameters (i.e., $g_j = s_j = .30$), but if we consider GDI, A-CDM estimates will be less discriminative than DINA and DINO ones. GDI is a more complex item discrimination index than can explain the differences in performance among the different models that were found in this study. The concept of item discrimination in CDM should be probability revisited.

A few limitations of this study are worth mentioning. To keep the scope of this study manageable, a few simplifications about factors affecting the CD-CAT performance were made. These included fixing the number of attributes, using a single method in estimating the attribute vectors, and focusing on the

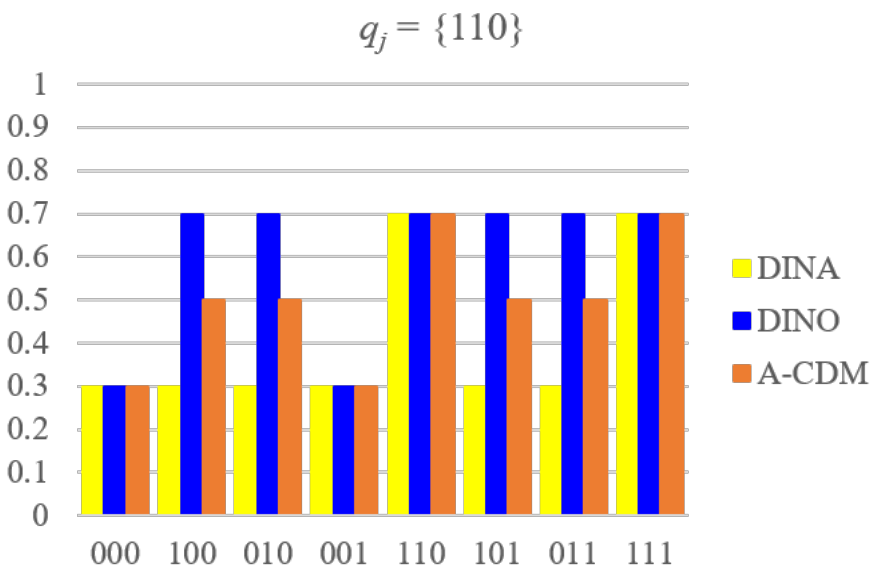

Figure 5.3: Item parameters for a three-attribute item for the three reduced CDMs.

unconstrained CD-CAT where neither exposure control nor content balancing were considered. The reason for that is that usually CDM applications are relatively low-stakes and, accordingly, test security is not a big concern. But if the test is high-stakes, exposure control becomes necessary. This will be considered in future research as well as different situations such as variable-length CD-CAT. In the data generation process, there was no reason to consider any particular attribute joint distribution. Therefore, latent classes were sampled from a uniform distribution. This favors the item bank calibrations. Different studies might explore the effect of the attribute joint distribution assuming a particular prior. Finally, this study focuses on what has come to be called low item quality in previous simulations studies (e.g., Ma et al., 2016; Sorrel et al., 2017a,b) because a higher accuracy improvement was expected. However, these values for item quality or discrimination estimates have been found to be in this range in many empirical applications outside educational measurement (de la Torre et al., 2017; Liu et al., 2013; Sorrel et al., 2016; Templin and Henson, 2006).

Chapter 6

General Discussion

In the last decades, cognitive diagnosis modeling (CDM) has emerged as a new framework for psychometric testing. Within this new framework, the underlying factors affecting performance on the test items are assumed to be discrete. Respondents are described using latent classes or profiles, rather than scores on a set of continuous latent variables. A new set of statistical models is used to estimate these latent profiles. These models are restricted latent class models referred to as cognitive diagnosis models (CDMs). This area of research is still in its early stages. As such, the current dissertation used empirical data and Monte Carlo methods to put forward in three directions within the CDM framework: broadening the area of application of CDMs, evaluating item-level model fit statistics, and introducing model comparison as a way of improving adaptive testing applications.

Traditionally, CDMs have been applied in educational measurement (Leighton and Gierl, 2007; Nichols et al., 2012) to classify students as masters and nonmasters of set of predetermined attributes (e.g., knowledge, skills, cognitive processes). These models were later applied for diagnosing psychological disorders (de la Torre et al., 2017; Templin and Henson, 2006). To a certain extent, CDMs can be understood as an extension of traditional multidimensional item response theory (IRT) and confirmatory factor analysis (CFA) models that is particularly suitable for modeling complex loading structures (Rupp and Templin, 2008). These complex loadings structures are characterized by within-item dimensionality (Adams et al., 1997). Within-item dimensionality, as opposed to between-item dimensionality, represents a situation where the performance on a specific item is due to multiple dimensions. Thus, there are several items measuring more than one dimension at a time. In this sense, CDM can be useful in other areas where this type of complex loading structure is presented. One of these areas is competency modeling in the area of Industrial-Organizational psychology, where typically situational judgment tests (SJTs) are used to measure multiple skills. Accordingly, one of the specific goals of this dissertation consisted in introducing CDM for the evaluation of SJT data.

The rest of the specific goals of the dissertation were related to model fit evaluation. The evaluation of

model fit is considered a crucial step in statistical modeling in general. In accordance to the importance of the model fit evaluation, considerable research has been undertaken in the past years within and without the CDM framework (e.g., Garrido et al. 2016; Hu et al. 2016; Huggins-Manley and Han 2017; Lei and Li 2016; Maydeu-Olivares et al. 2017; Sen and Bradshaw 2017). Most of this past research in the area of CDM have evaluated model fit at the test level. In contrast, the evaluation of item-level model fit statistics has received much less attention. There are many ways in which item-level fit evaluation can be useful. For example, absolute item-level fit evaluation can complement the analyses of the overall model fit. The sources of the overall misfit can be located using a lower level of analysis statistics. Traditionally, item- and item-pair levels have been considered (Chen and Thissen, 1997; Orlando and Thissen, 2000, 2003). Particularly, item-level fit evaluation can provide guidelines to practitioners on how to refine an instrument. In addition, implementing relative item-level fit indices allows conducting model comparison at the item level. Thus, the most appropriate model can be estimated for each item. This is of major relevance in CDM given the wide variety of models available and the empirical studies revealing that no one model can be deemed appropriate for all the test items (see, e.g., de la Torre et al. 2017; de la Torre and Lee 2013; Ravand 2016).

As will be discussed below, the results of the first item fit study indicated that the likelihood ratio (LR) test was the best performing statistic (Sorrel et al., 2017a). The main disadvantage of this statistics is that it requires estimating several combination of models, leading to a substantially long computation time. By contrast, a different statistic available, the Wald (W) test, only requires estimating the more general model. Following this, a new study was designed to introduce an approximation of the LR test incorporating this desirable feature of the W test. The new statistic was referred to as *two-step LR* (2LR) test because it is based on a two-step estimation approach (Sorrel et al., 2017b).

The next and last study of this dissertation evaluated whether item-level model comparison indices, such as the 2LR test, could be used to improve results in adapting testing. Considering how the test length affects the power of the performance of the item-level statistics (Sorrel et al., 2017a,b), the 2LR test was expected to perform very well under the usual item bank calibration conditions. The idea was to evaluate whether these indices could help researches to deal with poor quality data (e.g., small sample size, low item quality). Accordingly, this fourth study focused on low item quality and evaluated different sample size conditions.

This dissertation explores different factors such as sample size, test length, item quality, items complexity, and selection of an appropriate statistical model. These are some of the essential variables that must be considered at any test development and assessment processes. As such, some of the conclusions and directions provided in this dissertation are generalizable to other psychometric frameworks (e.g., factor analysis, structural equation modeling, IRT). A brief summary of the most relevant findings from

each study is presented next.

6.1 Most important findings from the Studies

6.1.1 Findings from Study 1: Application of cognitive diagnosis modeling to situational judgment tests data

Study 1 introduced the CDM approach to assess validity and reliability of SJT scores. Competences are typically evaluated using SJTs (McDaniel et al., 2001; Whetzel and McDaniel, 2009), which tend to be multidimensional even at the item level (Schmitt and Chan, 2006). Conventional methods for assessing the validity and reliability of SJT scores are based on classical test theory (CTT) (Christian et al., 2010; Ployhart and Weekley, 2006). This approach is not adequate considering SJT dimensionality. In a certain way, despite the promising criterion-related validity results of SJTs (see the meta-analyses of McDaniel et al., 2001; Mcdaniel et al., 2007), reliance on CTT has hampered further progress on different issues. For example, it is recognized that "there has been little success in understanding what SJTs really measure" (Ployhart and Weekley, 2006, p. 346). Being that so, *Study 1* illustrated the advantages of CDM over the traditional approach based on CTT using empirical data.

Four main advantages of the CDM approach were identified in the paper. First, this approach allows for a better understanding of the underlying internal structure of the SJT. In the empirical example, an initial list of four attributes was identified using prior research and theory: study habits, study attitudes, helping others, and generalized compliance. Expert ratings were used to develop an initial Q-matrix. Importantly, an empirical Q-matrix validation procedure was used to verify experts' decisions. The attributes were generally positively correlated. Second, the CDM approach can be used to explore what is the cognitive model that test takers engage when responding the items. Model fit evaluation provides information about how the attributes interact. In the empirical example, it was found that constraining the model to be conjunctive or disjunctive for all the items led to a significant loss of fit. A general model was then retained. However, it must be noted that some items seemed to follow a conjunctive process where all the attributes being measured by the item are required in order to have a high probability of success. In the same way, for other items, the mastery of one or more attributes could make up for lack of mastery in other attributes. This emphasises the importance of evaluating model fit at the item level, which is the focus of the rest of the studies included in this dissertation. Third, the CDM approach reveals why test scores relate to relevant criteria. Within the traditional approach, the lack of construct-level information make it hard to interpret the validity coefficients (Christian et al., 2010). In the empirical example, study habits was highly correlated with the grade point average and conscientiousness, and these correlation coefficients were somewhat higher than those estimates for the SJT sum score. Thus, most of the predictive power of the

SJT scores was due to this single attribute. Finally, the CDM approach introduces a new way of computing reliability. This is important because internal consistency of the SJT scores have been traditionally low (Catano et al., 2012), probably because alpha coefficient is not addecuate when items are hetegeneous (Miller, 1995). In the empirical example, reliability was assessed from a different angle. Results indicated that the classification accuracy of the four attributes was considerable high. Additionally, attributes scores define respondents strengths and weaknesses. This information can be usefully used in personnel selection and training programs (Weekley et al., 2015).

Overall, it is concluded that CDMs include a greater wealth of information in analyzing SJTs than traditional procedures based on CTT do. These advantages would depend on how carefully the initial list of attributes is developed. Practitioners can rely on prior research, theory, job analytic information, and think-aloud protocols. Whenever possible, the test should be designed from the very beginning (Tjoe and de la Torre, 2014).

6.1.2 Findings from Study 2: Inferential item fit evaluation in cognitive diagnosis modeling

Study 2 examined the performance of inferential item-level fit indices using Monte Carlo Methods. Of the many item-level model fit statistics that have been proposed in the literature, four inferential statistics were considered. The $S - X^2$ statistic introduced by (Orlando and Thissen, 2000, 2003) was selected because it has been studied extensively in the context of traditional IRT and has emerged as one of the most used in empirical applications (Amtmann et al., 2010; Glas and Falcón, 2003; Kang and Chen, 2008; Nieto et al., 2017). To keep the scope of the study manageable, different χ^2-like statistics such as Q_1 (Yen, 1981) were not considered. Although Q_1 has been previously used in the area of CDM (Sinharay and Almond, 2007; Wang et al., 2015), $S - X^2$ emerged as an alternative to Q_1 that addresses its main limitation, that is, the fact that the observed frequencies in the computation of Q_1 rely on the trait level estimates.

Within the area of item-level relative fit evaluation, the performance of the W test was compared to that of the other two classic methods, namely the LR and the Lagrange multiplier (LM) tests (Buse, 1982). Only the W test had been previously evaluated in CDM (de la Torre and Lee, 2013; Ma et al., 2016). In addition, data from current empirical CDM applications suggested that there might be differences in item discrimination regarding the constructs being assessed. Specifically, the discrimination estimates had been found to be lower in applications outside educational measurement (de la Torre et al., 2017; Sorrel et al., 2016; Templin and Henson, 2006). A lower item discrimination might be an expected result when CDMs are retrofitted, as well as poorer model fit (Rupp and Templin, 2008). It was still not clear how item discrimination affects the performance of the W test for some of the most commonly encountered

CDMs, namely the deterministic input, noisy "or" gate (DINA; Haertel, 1989) model, the deterministic inputs, noisy “or” gate (DINO; Templin and Henson, 2006) model, and the additive CDM (A-CDM, de la Torre, 2011). On the other hand, there might be reasons to prefer the LM test. The LM test tests for improvement of model fit, and only requires estimating the reduced CDM. Thus, the LM test holds promising in detecting the correct CDM when the general model is difficult to estimate.

Regarding absolute fit, $S - X^2$ was found to a satisfactory Type I error across all the simulated conditions. However, its power was far from reaching acceptable values. Regarding the relative fit statistic, overall comparisons favored LR and W tests over the LM test. Unfortunately, Type I error rates were only acceptable under high item quality conditions, although there were some notable exceptions. Particularly in the case of the LR test, the negative effect of item quality could be ameliorated by an increase in sample size and test length. Taking into account the high Type I error rates in some of the conditions, we can obtain the distribution of the statistics under the null hypothesis. We found that power was still generally high under medium to high item quality conditions. This bootstrap approximation is incorporated in the **ltm** (Rizopoulos, 2006) package and should be considered in the context of CDM. A different strategy would imply taking the best of each statistic, combining their results to make a decision. That is, among all the models that fit the data considering $S - X^2$, the one selected by LR or W test can be the one retained. The LR test was relatively more robust than the W test, but it had the limitation of being more computationally demanding. *Study 3* resumed this line for future research and introduced an efficient approximation to the LR test.

Overall, it is concluded that the statistics were generally not reliable under low item quality conditions due to a lack of power. Strategies to confront this problem include approximating the distribution of the statistics under the null hypothesis using a bootstrap approximation. In addition, $S - X^2$ can be used together with the LR or W test to make a well-founded considered decision. Either way, test model fit should be assessed as a whole and it needs to be ensured that the derived scores are valid and reliable.

6.1.3 Findings from Study 3: Proposal of an approximation to the likelihood ratio test

Study 3 introduced an efficient approximation to the LR test for item-level model comparison. Existing item-level model comparison analyses in the area of CDM were based on the W test (e.g., de la Torre et al., 2017; Ravand, 2016), which is the only test included in the software available (e.g., the **CDM** and **GDINA** packages in R; Ma and de la Torre, 2017; Robitzsch et al., 2017). According to the results from *Study 2*, both LR and W statistics did not perform well under poor quality data. This involved a small sample size, a short test length and, more especially, low item quality (Sorrel et al., 2017a). However, the LR test was relatively more robust than the W test. The main limitation of the LR test is its computational cost. Compared to the W test where only one model (i.e., the more general one) needs to be estimated, the

LR test requires estimating multiple combinations of models. In this study, the 2LR test was introduced as an efficient approximation to the LR test. This approximation is based on a two-step estimation procedure under the G-DINA model framework originally proposed by (de la Torre and Chen, 2011).

The simulation study results indicated that the performance of the 2LR and LR tests were very similar, being both of them preferred over the W test. Importantly, the computation of the 2LR test was remarkably faster. Regarding the manipulated factors, consistently with the previous study, item quality was found to have the greatest effect on the performance of the statistics. As expected, power decreased in the poor quality items conditions. It is noteworthy that 2LR power was the least affected in these conditions and tended to be high. Compared with *Study 2*, this study considered larger sample size (i.e., $N = 2,000$) and test lengths ($J = 60$). This was done in order to explore the extent to which these favorable conditions can compensate for a poor item quality. Indeed, this was the case for the LR and 2LR tests.

Overall, it is concluded that the 2LR test can be recommended for use in empirical research. Its performance will be acceptable provided that the item quality is medium or high. If the item quality is low, then a large sample size and test length are needed.

6.1.4 Findings from Study 4: Model selection in cognitive diagnosis computerized adaptive testing

Study 4 introduced item-level model comparisons indices as a way of improving cognitive diagnosis computerized adaptive testing (CD-CAT). Reduced CDMs have been usually preferred in empirical applications because they are easier to estimate and their parameters have a more straightforward interpretation, in comparison with general CDMs (e.g., Liu et al., 2013). Overall, though, this might not be a suitable approach given that reduced models make strong assumptions about the data. It is unlikely that this will be the case for all the items. This is more relevant in the context of adaptive testing because item banks tend to be considerably larger than standard versions of tests. A different strategy consists of estimating a general model for all items (e.g., Sorrel et al., 2018c). This should work fine provided the general model is estimated accurately. This will not be the case, for example, when the sample size is small or the number of parameters is high. Taking all of this into account, *Study 4* explored a different strategy. Specifically, the 2LR test (Sorrel et al., 2017b) developed in *Study 3* was used to select the best fitting model for each of the items in the item bank. The performance of CD-CATs based on these three strategies in terms of classification accuracy and item usage was explored. Manipulated factors included calibration sample size and Q-matrix complexity.

Results indicated that 2LR test improved the accuracy of the CD-CAT under all the simulated conditions. Accordingly, the same accuracy might be obtained with fewer items administered, a time saving of major importance in contexts (e.g., educational, medical) where testing time is always an issue.

Congruently with the effect of test length on the performance of this statistics (Sorrel et al., 2017a,b), we found that the true, generating reduced CDM was generally selected by the 2LR test. Regarding the manipulated factors, larger accuracy improvements were found when the calibration sample size was small and the Q-matrix was complex. On the other hand, CD-CATs based on a single reduced model led to a lower classification accuracy. This was ameliorated in a certain way when the calibration sample size was large and the Q-matrix had a simple structure, as in Liu et al. (2013). Importantly, even when a similar accuracy was obtained with the application of a single reduced model, items following a different reduced CDM were not selected by the adaptive procedure. This resulted in a poorer use of the item bank. Finally, procedures based on a general model led to optimal results provided the general model was accurately estimated (e.g., large sample size, simple Q-matrix). Otherwise, the classification accuracy was much lower, compared to the one based on the combination of models selected by 2LR.

Overall, it is concluded that item-level model selection indices such as the 2LR test can be a useful tool to improve classification accuracy and item usage in adaptive applications. These methodologies are very easy to implement using the software available (e.g., **CDM** and **GDINA** R packages).

6.2 Practical guidelines

Based on the results of the empirical and Monte Carlo studies, the following guidelines are proposed. First, interested researchers can consult the *Study 1* publication for a friendly introduction to the application of CDMs for validity and reliability assessment of SJT data. Different R packages were used, including the **CDM** (functions for cognitive diagnosis modeling) and **CTT** (a function for classical test theory analysis) (Willse, 2014) packages. At the moment of the publication of that paper, the code for the general method of Q-matrix validation (de la Torre and Chiu, 2016) was not available and had to be programmed in R. The entire code used in the paper is available for any interested reader by contacting the corresponding author of Sorrel et al. (2016). Note, however, that nowadays the **CDM** package is more complete and, besides, a new package with psychometric tools for CDM, the **GDINA** package, was released in April 13th, 2017. It is worth noting that the **GDINA** package includes two functions that can be of the interest of practitioners that are not familiar with the R programming environment. The $autoGDINA$ function conducts a series of CDM analyses automatically based on some user specifications: estimation of the G-DINA model, Q-matrix validation, item-level model selection, and final calibration. This sequence of analysis is congruent with the sequential steps in the application of CDMs described in Sorrel et al. (2016). The $startGDINA$ function starts a graphical user interface where all the analyses can be implemented without writing any R code. The package documentation can be consulted for additional details.

Second, the 2LR test is recommended for the determination of the most appropriate model for each item. The code is available by contacting the corresponding author of Sorrel et al. (2017b) and will be

published in an R package shortly. The main benefits of assessments based on multiple appropriate CDMs include a more straightforward interpretation of the item parameter estimates and an improvement in the classification accuracy (Ma et al., 2016; Rojas et al., 2012). These relative fit analyses should be complemented with absolute fit analysis. In that regard, the $S - X^2$ statistics has an adequate Type I error even under poor data conditions. Whenever data conditions are not the most optimal (i.e., small sample size, short test length, poor item quality), among all the models fitting the data considering $S - X^2$, the one flagged by the 2LR test can be retained.

Third, practitioners using CD-CAT methodologies are encouraged to use item-level model comparison statistics to select the most appropriate CDM for each item. This will have some advantages compared with a situation where the same single reduced CDM or a general CDM is fitted to all the items in the item bank. A CD-CAT based on the combination of models derived by the 2LR test will have a higher classification accuracy. This is especially so if the calibration sample size is small and the Q-matrix is complex, given that a general model will not be estimated accurately in that situations. Besides the classification accuracy improvement, a better use of the item pool will be obtained compared to the application of a single reduced CDM. Due to the lack of an R package for CD-CAT analyses, a new R code was created from the scratch. This code has already been employed in different publications (Sorrel et al., 2018c; Yigit et al., 2018). The program will be make available through an open-source package such as R.

6.3 Limitations and future lines of study

There are some limitations in this dissertation that should be noted. Most of them have been already mentioned in each specific paper. The most notable ones will be discussed here. One of these limitations is that the STJ used in *Study 1* was not originally developed within the CDM framework. This situation is referred to as *retrofitting* in the CDM literature (Liu et al., 2017; Gierl and Cui, 2008; Rupp and Templin, 2008). Some of the challenges and possibilities of retrofitting are discussed in Liu et al. (2017). Some of these problems include a poorer fit to the data and the lack of items measuring some specific attribute profiles. This might lead to a lower classification accuracy. Therefore, if a particular instrument is meant to be used for diagnostic purposes, a better approach would be considering that multidimensional structure from the very beginning. Fortunately, as noted by Liu et al. (2017), some of these problems such as the lack of fit or reliability can be investigated empirically. This was illustrated in *Study 1*.

Another limitation of this dissertation is that only four fit indices ($S - X^2$, W, LR, and LM) and one estimation method for these fit indices (MMLE-EM) were evaluated. Future studies may evaluate the performance of other indices and estimation methods, such as fit measures based on the residuals (Chen et al., 2013), and the root mean square error of approximation (RMSEA) and mean absolute difference

(MAD) indices (Henson et al., 2008; Kunina-Habenicht et al., 2012). In addition, there might be ways of improving the poor power results of the $S - X^2$ statistic. For example, Wang et al. (2015) applied the Q_1 item-fit statistic together with the method described in Stone (2000) for considering uncertainty in latent attribute estimation. This method dramatically increased the power rates of Q_1. Recently, Chalmers and Ng (2017) proposed a parametric bootstrap procedure that also considers trait level uncertainty. Future research should further explore these possibilities. On the other hand, the application of $S - X^2$ with sparse data and structurally incomplete test designs might be challenging. When the expected frequencies become small, the approximation to the χ^2 distribution deteriorates.

In addition, recent research detailed different ways of computing the item parameters standard errors in the context of CDM (Philipp et al., 2018). The W and LM tests computation requires using the estimated standard errors. It is then pivotal to explore whether the different ways of computing the standard errors affect the performance of these indices. Finally, the LM test can be also computed using the two-step estimation approach described in Sorrel et al. (2017b). The traditional implementation of the LM test can be compared with what would come to be called a two-step LM test. One would expect a better performance of the two-step LM test given that the attribute joint distribution will be more accurately estimated. The two-step approach might be also extended to other psychometric frameworks where models are nested, as is the case of IRT (e.g., the unidimensional logistic models).

Finally, an important limitation in the CD-CAT study is that data were generated using the traditional formulation of item discrimination, namely the item discrimination index (i.e., $IDI = P(\mathbf{1}) - P(\mathbf{0})$), but items were administrated based on the G-DINA model discrimination index (GDI; Kaplan et al., 2015). As it is discussed in *Study 4*, CD-CATs based on A-CDM were always worse than those based on the DINA and DINO models, even though DINA, DINO, and A-CDM were generated using the same item discrimination values. This indicates that probably the concept of item discrimination should be revisited. In addition, future studies might consider different items selection rules such as the mutual information (Wang, 2013) and large deviation (Liu et al., 2015) methods.

Appendix A

Contributed Work

This appendix contains a list of the contributed publications until the completion of this dissertation.

A.1 Main Author Contributions

- Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., and Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*, 19(3):506–532
  - Chapter 2
  - Published in February 2017 in *Organizational Research Methods*
- Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., and Barrada, J. R. (2017a). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement*, 41(8):614–631
  - Chapter 3
  - Published in May 2017 in *Applied Psychological Measurement*
- Sorrel, M. A., de la Torre, J., Abad, F. J., and Olea, J. (2017b). Two-step likelihood ratio test for item-level model comparison in cognitive diagnosis models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 13(S1):39
  - Chapter 4
  - Published in June 2017 in *Methodology*
- Sorrel, M. A., Abad, F. J., and Olea, J. (2018a). Model comparison as a way of improving cognitive diagnosis computerized adaptive testing. *Manuscript submitted for publication*
  - Chapter 5
  - Manuscript submitted for publication.

A.2 Related Research

- de la Torre, J. and Sorrel, M. A. (2017). Attribute classification accuracy and the monotonically constrained G-DINA model. In *The International Meeting of the Psychometric Society, Zurich, Switzerland, July 18-21, 2017*

- Sorrel, M. A., Barrada, J. R., and de la Torre, J. (2018b). Exchanging item selection rules from cognitive diagnosis modeling to traditional item response theory. *Manuscript submitted for publication*

- Sorrel, M. A., Yigit, H. D., and Kaplan, K. (2018c). CD-CAT implementation of the proportional reasoning assessment. *Manuscript submitted for publication*

- Yigit, H. D., Sorrel, M. A., and de la Torre, J. (2018). Computerized adaptive testing for multiple-choice cognitive diagnosis models. *Manuscript submitted for publication*

Appendix B

General Discussion (Spanish)

En las últimas décadas, el modelado diagnóstico cognitivo (MDC) se ha convertido en un nuevo marco para las pruebas psicométricas. Dentro de este nuevo marco, los factores subyacentes que afectan el rendimiento en los ítems en las pruebas se consideran discretos. Los examinados son descritos utilizando clases o perfiles latentes, en lugar de puntuaciones en un conjunto de variables latentes continuas. Se ha utilizado un nuevo conjunto de modelos estadísticos para estimar estos perfiles latentes. Estos modelos son modelos de clase latente confirmatorios denominados modelos de diagnóstico cognitivo (MDCs). Esta área de investigación todavía está en sus primeras etapas. Como tal, la presente tesis utilizó datos empíricos y métodos de simulación Monte Carlo para presentar avances en tres direcciones dentro del marco MDC: la ampliación del área de aplicación de los MDC, la evaluación de los estadísticos de ajuste a nivel de ítem y la propuesta del uso de estadísticos para la comparación de modelos como una forma de mejorar el función de las aplicaciones adaptativas.

Tradicionalmente, los MDC han sido aplicados en contextos de medición educativa (Leighton and Gierl, 2007; Nichols et al., 2012) para clasificar a los examinados como poseedores o no de un conjunto predeterminado de atributos (p.ej., conocimientos, habilidades, procesos cognitivos). Estos modelos se aplicaron posteriormente para diagnosticar trastornos psicológicos (de la Torre et al., 2017; Templin and Henson, 2006). En cierto sentido, los MDC se pueden entender como una extensión de los modelos tradicionales de teoría de respuesta al ítem multidimensional (TRIM) y análisis factorial confirmatorio (AFC) que es particularmente adecuada para modelar estructuras de pesos complejas (Rupp and Templin, 2008). Estas estructuras de pesos complejas se caracterizan por presentar dimensionalidad intra-ítem (Adams et al., 1997). La dimensionalidad intra-ítem, a diferencia de la dimensionalidad inter-ítem, representa una situación en la que el rendimiento de un ítem específico se debe a múltiples dimensiones. Por lo tanto, hay varios ítems que miden más de una dimensión a la vez. En este sentido, MDC puede ser útil en otras áreas donde este tipo de estructura de carga compleja es frecuente. Una de estas áreas es el modelado de competencias en el campo de la psicología industrial y organizacional, donde típicamente se

usan pruebas de juicio situacional (TJS) para medir multiples habilidades. En consecuencia, uno de los objetivos específicos de esta tesis consiste en la introducción de MDC para la evaluación de datos de TJS.

El resto de los objetivos específicos de la tesis están relacionados con la evaluación de ajuste del modelo. La evaluación del ajuste del modelo se considera un paso crucial en el modelado estadístico en general. De acuerdo con la importancia de la evaluación de ajuste del modelo, se han llevado a cabo investigaciones considerables en los últimos años dentro y fuera del marco del MDC (p.ej., Garrido et al. 2016; Hu et al. 2016; Huggins-Manley and Han 2017; Lei and Li 2016; Maydeu-Olivares et al. 2017; Sen and Bradshaw 2017). La mayor parte de esta investigación anterior en el área de MDC ha evaluado el ajuste del modelo a nivel del test. Por el contrario, la evaluación de las estadísticas de ajuste del modelo a nivel de ítem ha recibido mucha menos atención. Hay muchas formas en las que la evaluación de ajuste a nivel de ítem puede ser útil. Por ejemplo, la evaluación de ajuste absoluto del nivel de ítem puede complementar los análisis del ajuste global del modelo. Las fuentes del desajuste global se pueden ubicar utilizando estadísticos basados en un nivel más bajo de de análisis. Tradicionalmente, se han considerado los estadísticos a nivel de ítem ya nivel de pares de ítems (Chen and Thissen, 1997; Orlando and Thissen, 2000, 2003). En particular, la evaluación de ajuste a nivel de ítem puede proporcionar pautas a los profesionales sobre cómo refinar un instrumento. Además, la implementación de índices de ajuste relativos a nivel de ítem permite realizar la comparación de modelos a nivel de ítem. Por lo tanto, se puede estimar el modelo más apropiado para cada ítem. Esto es de gran relevancia en MDC dada la amplia variedad de modelos disponibles y los estudios empíricos que revelan que ningún modelo se puede considerar apropiado para todos los ítems de un test (vea, p.ej., de la Torre et al. 2017; de la Torre and Lee 2013; Ravand 2016).

Como se discutirá a continuación, los resultados del primer estudio de ajuste a nivel de ítem indicaron que la prueba de razón de verosimilitud (RV) fue el estadístico con mejor desempeño (Sorrel et al., 2017a). La principal desventaja de este estadístico es que requiere estimar varias combinaciones de modelos, lo que lleva a un tiempo de cálculo considerablemente largo. Por el contrario, un estadístico diferente disponible, la prueba de Wald (W), solo requiere estimar el modelo más general. Es por ello que, a continuación, se diseñó un nuevo estudio para introducir una aproximación de la prueba RV que incorpora esta característica deseable de la prueba W. La nueva prueba estadística se denominó *prueba de RV en dos pasos* (2RV) porque se basa en un enfoque de estimación en dos pasos (Sorrel et al., 2017b).

El siguiente y último estudio de esta tesis evaluó si los índices de comparación de modelos a nivel de ítems, como la prueba 2RV, podrían usarse para mejorar los resultados en aplicaciones adaptativas. Teniendo en cuenta cómo la longitud del test afecta a la potencia de los estadísticos a nivel de ítem (Sorrel et al., 2017a,b), se esperaba que la prueba 2RV funcionara muy bien en las condiciones habituales de calibración de bancos de ítems. La idea era evaluar si estos índices podrían ayudar a las investigaciones

a lidiar con datos de mala calidad (p.ej., tamaño muestral pequeño, baja calidad de los ítems). En consecuencia, este cuarto estudio se centró en condiciones de baja calidad de los ítems y evaluó diferentes condiciones de tamaño de muestra.

Esta tesis explora factores tales como el tamaño muestral, la longitud del test, la calidad de los ítems y su complejidad y la selección de un modelo estadístico apropiado. Estas son algunas de las variables esenciales que deben considerarse en cualquier proceso de evaluación y desarrollo de pruebas. Siendo así, algunas de las conclusiones y recomendaciones proporcionadas en esta tesis son generalizables a otros marcos psicométricos (por ejemplo, análisis de factores, modelado de ecuaciones estructurales, TRI). A continuación se presenta un breve resumen de los hallazgos más relevantes de cada estudio.

B.1 Los hallazgos más importantes de los estudios

B.1.1 Hallazgos del Estudio 1: Aplicación del modelado diagnostico cognitivo a test de juicio situacional

El *Estudio 1* introdujo el enfoque MDC para evaluar la validez y la fiabilidad de las puntuaciones de TJS. Las competencias se evalúan generalmente utilizando TJS (McDaniel et al., 2001; Whetzel and McDaniel, 2009), que tienden a ser multidimensionales incluso a nivel de ítem (Schmitt and Chan, 2006). Los métodos convencionales para evaluar la validez y la fiabilidad de las puntuaciones de TJS se basan en la teoría clásica de los tests (TCT) (Christian et al., 2010; Ployhart and Weekley, 2006). Este enfoque no es adecuado teniendo en cuenta la dimensionalidad de los TJS. En cierto sentido, a pesar de los prometedores resultados de validez referida a criterio que se ha obtenido con TJS (veánse los metanálisis de McDaniel et al., 2001; Mcdaniel et al., 2007), la confianza en la TCT ha obstaculizado el progreso en otros aspectos. Por ejemplo, se reconoce que "ha habido poco éxito en la comprensión de lo que los TJS realmente miden" (Ployhart and Weekley, 2006, p. 346). Siendo así, el *Estudio 1* ilustró las ventajas del enfoque basado en MDC sobre el enfoque tradicional basado en TCT usando datos empíricos.

En el artículo se identificaron cuatro principales ventajas del enfoque MDC. En primer lugar, este enfoque permite una mejor comprensión de la estructura interna subyacente del SJT. En el ejemplo empírico, se identificó una lista inicial de cuatro atributos utilizando la investigación y teoría disponibles: hábitos de estudio, actitudes hacia el estudio, ayudar a los demás y cumplimiento generalizado. Se utilizaron los juicios de expertos para desarrollar una matriz Q inicial. Es importante destacar que además se utilizó un procedimiento empírico de validación de la matriz Q para verificar las decisiones de los expertos. Los atributos generalmente correlacionaron positivamente. En segundo lugar, el enfoque MDC se puede utilizar para explorar cuál es el modelo cognitivo presente en los examinados cuando responden a los ítems. La evaluación de ajuste del modelo proporciona información sobre cómo interactúan los

atributos. En el ejemplo empírico, se encontró que restringir el modelo para que sea conjuntivo o disyuntivo para todos los ítems provocó una pérdida de ajuste significativa. Por tanto, se retuvo un modelo general. Sin embargo, debe tenerse en cuenta que algunos ítems parecían seguir un proceso conjuntivo en el que se requieren todos los atributos medidos por el ítem para tener una alta probabilidad de éxito. De la misma manera, para otros ítems, el dominio de uno o más atributos podría compensar la falta de dominio en otros atributos. Esto enfatiza la importancia de evaluar el ajuste del modelo a nivel de ítem. Este será el enfoque del resto de los estudios incluidos en esta tesis. En tercer lugar, el enfoque MDC revela por qué las puntuaciones de las pruebas se relacionan con los criterios relevantes. Dentro del enfoque tradicional, la falta de información a nivel de constructo dificulta la interpretación de los coeficientes de validez (Christian et al., 2010). En el ejemplo empírico, hábitos de estudio estaba altamente correlacionado con el promedio de calificaciones en la carrera y el rasgo de responsabilidad, y estos coeficientes de correlación fueron algo más altos que aquellos estimados para la puntuación suma total del test. Por lo tanto, la mayor parte del poder predictivo de las puntuaciones del TJS se debió a este único atributo. Finalmente, el enfoque MDC introduce una nueva forma de calcular la fiabilidad. Esto es importante porque la consistencia interna de las puntuaciones de TJS ha sido tradicionalmente baja (Catano et al., 2012), probablemente porque el coeficiente alfa no es adecuado cuando los ítems son heterogéneos en relación a la dimensionalidad. En el ejemplo empírico, la fiabilidad se evaluó desde un ángulo diferente. Los resultados indicaron que la precisión de las clasificaciones en los cuatro atributos fue considerablemente alta. Además, las puntuaciones en los atributos definen las fortalezas y debilidades de los examinados. Esta información puede ser utilizada en selección de personal y en programas de entrenamiento (Weekley et al., 2015).

En general, se concluye que los MDCs incluyen una mayor riqueza de información en el análisis de TJS que los procedimientos tradicionales basados en TCT. Estas ventajas dependerán de cuán cuidadosamente se desarrolle la lista inicial de atributos. Los profesionales pueden confiar en la investigación previa, la teoría, la información analítica del trabajo y el análisis de protocolos de pensar en voz alta. Siempre que sea posible, la prueba debe diseñarse desde el principio (Tjoe and de la Torre, 2014).

B.1.2 Hallazgos del Estudio 2: Ajuste inferencial a nivel de ítem en modelado diagnostico cognitivo

El *Estudio 2* examinó el funcionamiento de las medidas de ajuste inferencial a nivel de ítem utilizando un estudio de simulación Monte Carlo. De los estadísticos de ajuste a nivel de ítem que se han propuesto en la literatura, se consideraron cuatro estadísticos inferenciales. El estadístico $S - X^2$ propuesto por Orlando and Thissen (2000, 2003) fue seleccionado porque se ha estudiado extensamente en el contexto de la TRI tradicional y se ha convertido en uno de los más utilizados en aplicaciones empíricas (Amtmann

et al., 2010; Glas and Falcón, 2003; Kang and Chen, 2008; Nieto et al., 2017). A fin de que el ámbito del estudio no resultara demasiado extenso, otros estadísticos χ^2 como Q_1 (Yen, 1981) no se incluyeron. A pesar de que Q_1 ha sido empleado previamente en el área de los MDCs (Sinharay and Almond, 2007; Wang et al., 2015), $S-X^2$ surgió como una alternativa a Q_1 que aborda su principal limitación, es decir, el hecho de que las frecuencias observadas que se usan en el cálculo de Q_1 depende de los niveles de rasgo estimados.

Dentro del ámbito del ajuste relativo a nivel de ítem, el funcionamiento de la prueba W se comparó al de los otros dos métodos clásicos, la prueba RV y el test de multiplicadores de Lagrange (LM) (Buse, 1982). El test W había sido el único evaluado previamente en MDC (de la Torre and Lee, 2013; Ma et al., 2016). Además, los datos obtenidos en las aplicaciones empíricas de MDCs sugerían posibles diferencias en la discriminación de los ítems en función de los constructos medidos. En particular, se había encontrado que las estimaciones de discriminación eran menores en aplicaciones fuera de la medición educativa (de la Torre et al., 2017; Sorrel et al., 2016; Templin and Henson, 2006). Una discriminación más baja de los ítem podría ser un resultado esperado cuando los MDC se retroadaptan, así como un ajuste más pobre del modelo a los datos (Rupp and Templin, 2008). Todavía no estaba claro cómo la discriminación de los ítems afecta el rendimiento de la prueba W para algunos de los MDCs más comúnmente encontrados, como son los modelos deterministic input, noisy "or" gate (DINA; Haertel, 1989), deterministic inputs, noisy “or” gate (DINO; Templin and Henson, 2006) y el additive CDM (A-CDM, de la Torre, 2011). Por otro lado, existen motivos para preferir la prueba LM. La prueba LM evalúa la mejora en ajuste del modelo, y sólo requiere estimar el MDC reducido. Por lo tanto, la prueba LM es prometedora para detectar el MDC correcto cuando el modelo general es difícil de estimar

Con respecto al ajuste absoluto, se encontró que $S-X^2$ tenía un error Tipo I satisfactorio en todas las condiciones simuladas. Sin embargo, su potencia estadística estaba lejos de alcanzar valores aceptables. Con respecto a los estadísticos de ajuste relativo, las comparaciones en general favorecieron las pruebas RV y W sobre la prueba LM. Desafortunadamente, las tasas de error de Tipo I sólo fueron aceptables en condiciones de alta calidad de los ítems, aunque hubo algunas excepciones notables. Particularmente en el caso de la prueba RV, el efecto negativo de la calidad de los ítems podía mejorarse mediante un aumento en el tamaño de la muestra y la longitud del test. Teniendo en cuenta las altas tasas de error de Tipo I en algunas de las condiciones, se obtuvo la distribución de los estadísticos bajo la hipótesis nula. Descubrimos que la potencia estadística todavía era alta generalmente en condiciones de media a alta calidad de los ítems. Esta aproximación de remuestreo está incorporada en el paquete **ltm** (Rizopoulos, 2006) y debe considerarse en el contexto de MDC. Una estrategia diferente implicaría tomar lo mejor de cada estadístico, combinando sus resultados para tomar una decisión. Es decir, entre todos los modelos que se ajustan a los datos considerando $S-X^2$, el que se seleccionó mediante la prueba RV o W puede ser

el que se conserve. La prueba RV fue relativamente más robusta que la prueba W, pero tenía la limitación de ser más exigente desde el punto de vista computacional. El *Estudio 3* reanudó esta línea para futuras investigaciones e introdujo una aproximación eficiente a la prueba RV.

En general, se concluye que los estadísticos generalmente no presentan un funcionamiento adecuado en condiciones de baja calidad de los ítems debido a la falta de potencia estadística. Las estrategias para enfrentar este problema incluyen aproximar la distribución de las estadísticas bajo la hipótesis nula utilizando métodos de remuestreo. Además, $S - X^2$ se pueden usar junto con la prueba RV o W para tomar una decisión bien fundada. De cualquier forma, el ajuste debe evaluarse como un todo y se debe garantizar que las puntuaciones derivadas sean válidas y fiables.

B.1.3 Hallazgos del Estudio 3: Propuesta una aproximación a la prueba de razón de verosimilitud

El *Estudio 3* introdujo una aproximación eficiente a la prueba RV para la comparación de modelos a nivel de ítem. Los análisis de comparación de modelos a nivel de ítem existentes en el área de MDC se basaron en la prueba W (p.ej., de la Torre et al., 2017; Ravand, 2016), que es la única incluida en el software disponible (p.ej., los paquetes de R **MDC** y **GDINA**; Ma and de la Torre, 2017; Robitzsch et al., 2017). De acuerdo con los resultados del *Estudio 2*, tanto los estadísticos RV como W no tuvieron un buen desempeño con datos de baja calidad. Esto involucró un tamaño muestral pequeño, una longitud del test corta y, más especialmente, una calidad baja de los ítems. Sin embargo, la prueba de RV fue relativamente más robusta que la prueba W. La principal limitación de la prueba RV es su coste computacional. En comparación con la prueba W, donde sólo se necesita estimar un modelo (es decir, el más general), la prueba RV requiere estimar múltiples combinaciones de modelos. En este estudio, la prueba 2RV se introdujo como una aproximación eficiente a la prueba RV. Esta aproximación se basa en un procedimiento de estimación en dos pasos en el marco del modelo G-DINA originalmente propuesto por de la Torre and Chen (2011).

En general, se concluye que la prueba 2RV se puede recomendar para su uso en investigación empírica. Su rendimiento será aceptable siempre que la calidad de los ítems sea media o alta. Si es baja, entonces se requieren un mayor tamaño muestral y una mayor longitud del test.

B.1.4 Hallazgos del Estudio 4: Selección de modelos en test adaptativos informatizados de diagnóstico cognitivo

El *Estudio 4* introdujo los índices de comparaciones de modelos a nivel de ítems como una forma de mejorar el funcionamiento de los test adaptativos informatizados de diagnóstico cognitivo (TAI-DC). Los MDC reducidos generalmente se han preferido en aplicaciones empíricas porque son más fáciles de

estimar y sus parámetros tienen una interpretación más directa en comparación con los MDCs generales (p.ej., Liu et al., 2013). En general, sin embargo, este podría no ser un enfoque adecuado, dado que los modelos reducidos hacen suposiciones fuertes sobre los datos, por lo que podrían no ser apropiados para todos los ítems. Esto es más relevante en el contexto de pruebas adaptativas porque los bancos de ítems tienden a ser bastante grandes. Una estrategia diferente consiste en estimar un modelo general para todos los ítems (por ejemplo, Sorrel et al., 2018c). Esto debería funcionar bien siempre que el modelo general se calcule con precisión. Este no será el caso, por ejemplo, cuando el tamaño de la muestra sea pequeño o la cantidad de parámetros sea alta. Tomando todo esto en cuenta, el *Estudio 4* exploró una estrategia diferente. Específicamente, la prueba 2RV (Sorrel et al., 2017b) desarrollada en el *Estudio 3* se usó para seleccionar el mejor modelo para cada uno de los ítems presentes en el banco de ítems. Se exploró el rendimiento de los TAI-DC basados en estas tres estrategias en términos de precisión de clasificación y uso de los ítems. Los factores manipulados incluyeron el tamaño de la muestra de calibración y la complejidad de la matriz Q.

Los resultados indicaron que la prueba 2RV mejoró la precisión del TAI-DC en todas las condiciones simuladas. En consecuencia, la misma precisión podría obtenerse con menos ítems administrados, un ahorro de tiempo de gran importancia en contextos donde el tiempo de evaluación es siempre un problema (p.ej., educativos, médicos). De manera congruente con el efecto de la longitud del test en el rendimiento de este estadístico (Sorrel et al., 2017a,b), encontramos que el MDC verdadero que se usó para generar los datos era generalmente seleccionado por la prueba 2RV. Con respecto a los factores manipulados, las mejoras en precisión fueron mayores cuando el tamaño de la muestra de calibración era pequeño y la matriz Q era compleja. Por otro lado, los TAI-DC basados en un único modelo reducido condujeron a una menor precisión de las clasificaciones. Esta precisión mejoró en cierta manera cuando el tamaño de la muestra de calibración era grande y la matriz Q tenía una estructura simple, como en Liu et al. (2013). Es importante destacar que, incluso cuando se obtuvo una precisión similar con la aplicación de un único modelo reducido, los ítems que fueron generados con un MDC reducido diferente no fueron selecionados por el algoritmo adaptativo. Esto resultó en un uso más pobre del banco de ítems. Finalmente, los procedimientos basados en un modelo general condujeron a resultados óptimos siempre que el modelo general se estimó con precisión (p.ej., tamaño muestral grande, matriz Q simple). En caso contrario, la precisión de la clasificación fue mucho menor en comparación con la basada en la combinación de modelos seleccionados por 2RV.

En general, se concluye que los índices de selección de modelos a nivel de ítem, como la prueba 2RV, pueden ser una herramienta útil para mejorar la precisión de las clasificaciones y el uso de los ítems en aplicaciones adaptativas. Estas metodologías son muy fáciles de implementar utilizando el software disponible (p.ej., **CDM** and **GDINA** R packages).

B.2 Guías prácticas

En base en los resultados de los estudios empíricos y de simulación Monte Carlo, se proponen las siguientes pautas. Primero, los investigadores interesados pueden consultar la publicación del *Estudio 1* para una introducción amistosa a la aplicación de los MDC para la evaluación de validez y fiabilidad de los datos obtenidos con SJT. Se utilizaron diferentes paquetes R, incluidos los paquetes **CDM** (funciones para el modelado de diagnóstico cognitivo) y **CTT** (una función para el análisis de teoría clásica) (Willse, 2014). En el momento de la publicación de ese documento, el código para el método general de validación Q matrix (de la Torre and Chiu, 2016) no estaba disponible y tuvo que ser programado en R. El código completo utilizado en el documento está disponible para cualquier lector interesado poniéndose en contacto con el autor de correspondencia de Sorrel et al. (2016). Sin embargo, debe tenerse en cuenta que hoy en día el paquete **CDM** es más completo y que, además, un nuevo paquete con herramientas psicométricas para MDC, el paquete **GDINA** fue publicado el 13 de abril de 2017. Vale la pena señalar que el paquete **GDINA** incluye dos funciones que pueden ser de interés para los profesionales que no están familiarizados con el entorno de programación R. La función $autoGDINA$ realiza una serie de análisis MDC de forma automática en función de algunas especificaciones del usuario: estimación del modelo G-DINA, validación de la matriz Q, selección del modelo a nivel de ítem y calibración final. Esta secuencia de análisis es congruente con los pasos secuenciales en la aplicación de los MDC descritos en Sorrel et al. (2016). La función $startGDINA$ inicia una interfaz gráfica de usuario donde todos los análisis pueden implementarse sin escribir ningún código en R. La documentación del paquete se puede consultar para obtener detalles adicionales.

En segundo lugar, se recomienda la prueba 2RV para determinar el modelo más apropiado para cada ítem. El código está disponible al contactar al autor correspondiente de Sorrel et al. (2017b) y se publicará en un paquete R en breve. Los principales beneficios de las evaluaciones basadas en múltiples MDC apropiados incluyen una interpretación más directa de las estimaciones de parámetros de los ítems y una mejora en la precisión de las clasificaciones (de la Torre and Sorrel, 2017; Ma et al., 2016; Rojas et al., 2012). Estos análisis de ajuste relativo deben complementarse con un análisis de ajuste absoluto. En ese sentido, el estadístico $S - X^2$ presenta una tasa de error tipo I adecuada incluso en condiciones de datos deficientes. Siempre que las condiciones de datos no sean las más óptimas (es decir, tamaño muestral pequeño, longitud del corta, calidad los ítems deficiente), entre todos los modelos que se ajustan correctamente a los datos considerando $S - X^2$, se puede retener el señalado por la prueba 2RV.

En tercer lugar, se recomienda a los profesionales que usan metodologías de TAI-DC utilizar estadísticos de comparación de modelos a nivel de ítem para seleccionar el MDC más apropiado para cada ítem. Esto tendrá algunas ventajas en comparación con una situación donde un único MDC reducido o un MDC general se ajusta a todos los ítems presentes en el banco. Un TAI-DC basado en la combinación

de modelos derivados por la prueba 2RV tendrá una mayor precisión de las clasificaciones. Esto es especialmente cierto si el tamaño de muestra de calibración es pequeño y la matriz Q es compleja, dado que un modelo general no se estimará con precisión en esas situaciones. Además de la mejora en la precisión de las clasificaciones, se obtendrá un mejor uso de los ítems en comparación con la aplicación de un único MDC reducido. Debido a la falta de un paquete R para los análisis TAI-DC, se creó un nuevo código R desde cero. Este código ya se ha utilizado en diferentes publicaciones (Sorrel et al., 2018c; Yigit et al., 2018). El programa estará disponible a través de un paquete de código abierto como R.

B.3 Limitaciones y futuras lineas de estudio

Hay algunas limitaciones en esta disertación que deben tenerse en cuenta. La mayoría de estas limitaciones ya han sido mencionadas en cada artículo específico. Sólo las más notables son discutidas aquí. Una de estas limitaciones es que el TJS utilizado en el *Estudio 1* no se desarrolló originalmente dentro del marco MDC. Esta situación se conoce como *retroadaptación* en la literatura de MDC (Liu et al., 2017; Gierl and Cui, 2008; Rupp and Templin, 2008). Algunos de los desafíos y posibilidades de la retroadaptación se discuten en Liu et al. (2017). Algunos de estos problemas incluyen un ajuste más pobre a los datos y la falta de ítems que midan algunos perfiles de atributos específicos. Esto podría conducir a una menor precisión de las clasificaciones. Por lo tanto, si un instrumento en particular está destinado a ser utilizado con fines de diagnóstico, un mejor enfoque sería considerar esa estructura multidimensional desde el principio. Afortunadamente, como se señala en Liu et al. (2017), algunos de estos problemas, como la falta de ajuste o fiabilidad, se pueden investigar empíricamente. Esto se ilustra en el *Estudio 1*.

Otra limitación de esta disertación es que sólo se evaluaron cuatro índices de ajuste ($S-X^2$, W, RV y LM) y un método de estimación para estos índices de ajuste (MMLE-EM). Los estudios futuros pueden evaluar el rendimiento de otros índices y métodos de estimación, como medidas de ajuste basadas en los residuales (Chen et al., 2013) y otras medidas descriptivas como el error cuadrático medio de aproximación (RMSEA) y la diferencia absoluta media (MAD) (Henson et al., 2008; Kunina-Habenicht et al., 2012). Además, podría haber formas de mejorar los pobres resultados de potencia estadística encontrados para $S-X^2$. Por ejemplo, Wang et al. (2015) aplicó el estadístico Q_1 junto con el método descrito en Stone (2000) para considerar la incertidumbre en la estimación de los atributos latentes. Este método aumentó drásticamente la potencia estadística de Q_1. Recientemente, Chalmers and Ng (2017) propuso un método paramétrico de remuestreo que también considera la incertidumbre del nivel de rasgo. La investigación futura debería explorar más a fondo estas posibilidades. Por otro lado, la aplicación de $S-X^2$ con datos dispersos y diseños de prueba estructuralmente incompletos podría ser un desafío. Cuando las frecuencias esperadas se vuelven pequeñas, la aproximación a la distribución χ^2 se deteriora.

Además, investigaciones recientes detallaron diferentes formas de calcular los errores típicos de los

parámetros de los ítems en el contexto de MDC (Philipp et al., 2018). El cálculo de las pruebas W y LM requiere el uso de los errores típicos estimados. Entonces, es fundamental explorar si las diferentes formas de calcular los errores típicos afectan el rendimiento de estos índices. Finalmente, la prueba LM también se puede calcular utilizando el enfoque de estimación en dos pasos que se describe en Sorrel et al. (2017b). La implementación tradicional de la prueba LM se puede comparar con lo que se vendría a llamar una prueba LM en dos pasos. Podría esperarse un mejor funcionamiento de la prueba de LM en dos pasos dado que la distribución conjunta de atributos se estimará con mayor precisión. El enfoque en dos pasos también podría extenderse a otros marcos psicométricos donde los modelos están anidados, como es el caso del TRI (p.ej., los modelos logísticos unidimensionales).

Para terminar, una limitación importante en el estudio TAI-DC es que los datos se generaron usando la formulación tradicional de discriminación de ítems: el denominado índice de discriminación de ítems (es decir, $IDI = P(\mathbf{1}) - P(\mathbf{0})$), pero los ítems fueron administrados en base al índice de discriminación del modelo G-DINA (GDI; Kaplan et al., 2015). Tal y como se analiza en el *Estudio 4*, los TAI-DC basados en A-CDM siempre fueron peores que los basados en los modelos DINA y DINO, aunque DINA, DINO y A-CDM se generaron utilizando los mismos valores de discriminación de los ítems. Esto indica que probablemente el concepto de discriminación de los ítems debería revisarse. Además, los estudios futuros podrían considerar diferentes reglas de selección de ítems, como los métodos de información mutua (Wang, 2013) y gran desviación (Liu et al., 2015).

Bibliography

Adams, R. J., Wilson, M., and Wang, W.-c. (1997). The multidimensional random coefficients multinomial logit model. *Applied psychological measurement*, 21(1):1–23.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Akbay, L. and Kaplan, M. (2017). Transition to multidimensional and cognitive diagnosis adaptive testing: An overview of CAT. *The Online Journal of New Horizons in Education-January*, 7(1).

Amtmann, D., Cook, K. F., Jensen, M. P., Chen, W.-H., Choi, S., Revicki, D., Cella, D., Rothrock, N., Keefe, F., Callahan, L., et al. (2010). Development of a PROMIS item bank to measure pain interference. *Pain*, 150(1):173–182.

Baghaei, P. and Ravand, H. (2015). A cognitive processing model of reading comprehension in English as a foreign language using the linear logistic test model. *Learning and Individual Differences*, 43:100–105.

Bartram, D. (2005). The great eight competencies: A criterion-centric approach to validation. *Journal of applied psychology*, 90(6):1185.

Bley, S. (2017). Developing and validating a technology-based diagnostic assessment using the evidence-centered game design approach: An example of intrapreneurship competence. *Empirical Research in Vocational Education and Training*, 9(1):6.

Bradshaw, L., Izsák, A., Templin, J., and Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational measurement: Issues and practice*, 33(1):2–14.

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician*, 36(3a):153–157.

Catano, V. M., Brochu, A., and Lamerson, C. D. (2012). Assessing the reliability of situational judgment tests used in high-stakes situations. *International Journal of Selection and Assessment*, 20(3):333–346.

Chalmers, R. P. and Ng, V. (2017). Plausible-value imputation statistics for detecting item misfit. *Applied Psychological Measurement*, 41(5):372–387.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, 80(1):1–20.

Chang, H.-H. and Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20(3):213–229.

Chen, J., de la Torre, J., and Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, 50(2):123–140.

Chen, J. and Zhou, H. (2017). Test designs and modeling under the general nominal diagnosis model framework. *PloS one*, 12(6):e0180016.

Chen, W.-H. and Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3):265–289.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, 74(4):619.

Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: The modified maximum global discrimination index method. *Educational and Psychological Measurement*, 70(6):902–913.

Choi, K. M., Lee, Y.-S., and Park, Y. S. (2015). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science & Technology Education*, 11(6).

Christian, M. S., Edwards, B. D., and Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, 63(1):83–117.

Cui, Y. and Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement*, 46(4):429–449.

Cui, Y. and Li, J. (2015). Evaluating person fit for cognitive diagnostic assessment. *Applied Psychological Measurement*, 39(3):223–238.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of educational measurement*, 45(4):343–362.

de la Torre, J. (2009a). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, 33(3):163–183.

de la Torre, J. (2009b). DINA model and parameter estimation: A didactic. *Journal of educational and behavioral statistics*, 34(1):115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76(2):179–199.

de la Torre, J. and Chen, J. (2011). Estimating different reduced cognitive diagnosis models using a general framework. In *Annual Meeting of the National Council on Measurement in Education*.

de la Torre, J. and Chiu, C.-Y. (2016). General method of empirical Q-matrix validation. *Psychometrika*, 81(2):253–273.

de la Torre, J. and Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3):333–353.

de la Torre, J., Hong, Y., and Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement*, 47(2):227–249.

de la Torre, J. and Lee, Y.-S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, 50(4):355–373.

de la Torre, J. and Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, 20(2):89–97.

de la Torre, J. and Sorrel, M. A. (2017). Attribute classification accuracy and the monotonically constrained G-DINA model. In *The International Meeting of the Psychometric Society, Zurich, Switzerland, July 18-21, 2017*.

de la Torre, J., van der Ark, L. A., and Rossi, G. (2017). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, pages 1–16.

DiBello, L. V., Roussos, L. A., and Stout, W. (2007). A review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of statistics*, 26:979–1030.

DiBello, L. V., Stout, W. F., and Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. *Cognitively diagnostic assessment*, pages 361–389.

Doornik, J. A. and Ooms, M. (2007). *Introduction to Ox: An Object-Oriented Matrix Language.* London, England: Timberlake Consultants Ltd.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 56(3):495–515.

Feasel, K., Henson, R., and Jones, L. (2004). Analysis of the gambling research instrument (GRI). *Unpublished manuscript*.

Fischer, G. H. (1997). Unidimensional linear logistic rasch models. In *Handbook of modern item response theory*, pages 225–243. Springer.

García, P. E., Olea, J., and de la Torre, J. (2014). Application of cognitive diagnosis models to competency-based situational judgment tests. *Psicothema*, 26(3):372–377.

Garrido, L. E., Abad, F. J., and Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via monte carlo simulation. *Psychological methods*, 21(1):93.

Gierl, M. J. and Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement*, 6:263–268.

Glas, C. A. and Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, 27(2):87–106.

Gu, Y. and Xu, G. (2017). The sufficient and necessary condition for the identifiability and estimability of the DINA model. *arXiv preprint arXiv:1711.03174*.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26(4):301–321.

Hagenaars, J. A. and McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge University Press.

Hansen, M. (2013). *Hierarchical item response models for cognitive diagnosis*. PhD thesis, UCLA.

Hansen, M., Cai, L., Monroe, S., and Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology*, 69(3):225–252.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* PhD thesis, ProQuest Information & Learning.

Henson, R. and Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, 29(4):262–277.

Henson, R., Roussos, L., Douglas, J., and He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement*, 32(4):275–288.

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74(2):191.

Hsu, C.-L., Wang, W.-C., and Chen, S.-Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement*, 37(7):563–582.

Hu, J., Miller, M. D., Huggins-Manley, A. C., and Chen, Y.-H. (2016). Evaluation of model fit in cognitive diagnosis models. *International Journal of Testing*, 16(2):119–141.

Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment*, 15(3):1–7.

Huebner, A. and Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, 71(2):407–419.

Huggins-Manley, A. C. and Han, H. (2017). Assessing the sensitivity of weighted least squares model fit indexes to local dependence in item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(3):331–340.

Huo, Y. and de la Torre, J. (2014). Estimating a cognitive diagnostic model for multiple strategies via the EM algorithm. *Applied Psychological Measurement*, 38(6):464–485.

Junker, B. W. and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3):258–272.

Kang, T. and Chen, T. T. (2008). Performance of the generalized S-X2 item fit index for polytomous IRT models. *Journal of Educational Measurement*, 45(4):391–406.

Kaplan, M., de la Torre, J., and Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied psychological measurement*, 39(3):167–188.

Kunina-Habenicht, O., Rupp, A. A., and Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, 49(1):59–81.

Kuo, B.-C., Chen, C.-H., and de la Torre, J. (2017). A cognitive diagnosis model for identifying coexisting skills and misconceptions. *Applied Psychological Measurement*, page 0146621617722791.

Lee, Y.-S., Park, Y. S., and Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing*, 11(2):144–177.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Lei, P.-W. and Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and Q-matrices. *Applied Psychological Measurement*, 40(6):405–417.

Leighton, J. and Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Li, H., Hunter, C. V., and Lei, P.-W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, 33(3):391–409.

Li, H. and Suen, H. K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment*, 18(1):1–25.

Liu, H.-Y., You, X.-F., Wang, W.-Y., Ding, S.-L., and Chang, H.-H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of classification*, 30(2):152–172.

Liu, J., Ying, Z., and Zhang, S. (2015). A rate function approach to computerized adaptive testing for cognitive diagnosis. *Psychometrika*, 80(2):468–490.

Liu, R., Huggins-Manley, A. C., and Bulut, O. (2017). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement*, doi:0013164416685599.

Liu, Y., Douglas, J. A., and Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement*, 33(8):579–598.

Liu, Y., Tian, W., and Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics*, 41(1):3–26.

Ma, W. and de la Torre, J. (2017). GDINA: The generalized DINA model framework. *R package version 1.4.2. Available online at: http://CRAN. R-project. org/package= GDINA*.

Ma, W., Iaconangelo, C., and de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, 40(3):200–217.

Ma, W. and Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, 69(3):253–275.

Magis, D. and Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software*, 76(1):1–19.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64(2):187–212.

Maydeu-Olivares, A., Fairchild, A. J., and Hall, A. G. (2017). Goodness of fit in item factor analysis: Effect of the number of response alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(4):495–505.

Mcdaniel, M. A., Hartman, N. S., Whetzel, D. L., and Grubb, W. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel psychology*, 60(1):63–91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., and Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, 86(4):730.

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Lawrence Erlbaum.

McGlohen, M. and Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, 40(3):808–821.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modelling*, 2:255–273.

Millon, T., Millon, C., Davis, R., and Grossman, S. (2009). *MCMI-III Manual (4th ed.)*. Mineapolis, MN: Pearson Assessments.

Minchen, N. and de la Torre, J. (2016). The continuous G-DINA model and the Jensen-Shannon divergence. In *International Meeting of the Psychometric Society, Asheville, NC, July 11-15, 2016*.

Mislevy, R. J. (1994). Probability-based inference in cognitive diagnosis. *ETS Research Report Series*, 1994(1).

Muthén, L. and Muthén, B. (2013). *Mplus 7.11*. Los Angeles, CA: Muthén & Muthén.

Nichols, P. D., Chipman, S. F., and Brennan, R. L. (2012). *Cognitively diagnostic assessment*. Routledge.

Nieto, M. D., Abad, F. J., Hernández-Camacho, A., Garrido, L. E., Barrada, J. R., Aguado, D., and Olea, J. (2017). Calibrating a new item pool to adaptively assess the big five. *Psicothema*, 29(3).

Olea, J., Barrada, J. R., Abad, F. J., Ponsoda, V., and Cuevas, L. (2012). Computerized adaptive testing: The capitalization on chance problem. *The Spanish journal of psychology*, 15(1):424–441.

Orlando, M. and Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24(1):50–64.

Orlando, M. and Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, 27(4):289–298.

Philipp, M., Strobl, C., de la Torre, J., and Zeileis, A. (2018). On the estimation of standard errors in cognitive diagnosis models. *Journal of Educational and Behavioral Statistics*, doi:1076998617719728.

Ployhart, R. E. and Weekley, J. A. (2006). Situational judgment: Some suggestions for future science and practice. *Situational judgment tests: Theory, measurement, and application*, pages 345–350.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, 34(8):782–799.

Rizopoulos, D. (2006). ltm: An r package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5):1–25.

Robitzsch, A., Kiefer, T., George, A. C., and Uenlue, A. (2017). Cdm: Cognitive diagnosis modeling. *R package version*, 6.0-101.

Rojas, G., de la Torre, J., and Olea, J. (2012). Choosing between general and specific cognitive diagnosis models when the sample size is small. In *Annual Meeting of the National Council on Measurement in Education, Vancouver, Canada*.

Rupp, A. A. and Mislevy, R. J. (2007). Cognitive foundations of structured item response models.

Rupp, A. A., Templin, J., and Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Rupp, A. A. and Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, 6(4):219–262.

Schmitt, N. and Chan, D. (2006). Situational judgment tests: Method or construct. *Situational judgment tests: Theory, measurement, and application*, pages 135–155.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Sen, S. and Bradshaw, L. (2017). Comparison of relative fit indices for diagnostic model selection. *Applied Psychological Measurement*, 41(6):422–438.

Shute, V. J., Leighton, J. P., Jang, E. E., and Chu, M.-W. (2016). Advances in the science of assessment. *Educational Assessment*, 21(1):34–59.

Sinharay, S. and Almond, R. G. (2007). Assessing fit of cognitive diagnostic models a case study. *Educational and Psychological Measurement*, 67(2):239–257.

Sorrel, M. A., Abad, F. J., and Olea, J. (2018a). Model comparison as a way of improving cognitive diagnosis computerized adaptive testing. *Manuscript submitted for publication*.

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., and Barrada, J. R. (2017a). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement*, 41(8):614–631.

Sorrel, M. A., Barrada, J. R., and de la Torre, J. (2018b). Exchanging item selection rules from cognitive diagnosis modeling to traditional item response theory. *Manuscript submitted for publication*.

Sorrel, M. A., de la Torre, J., Abad, F. J., and Olea, J. (2017b). Two-step likelihood ratio test for item-level model comparison in cognitive diagnosis models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 13(S1):39.

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., and Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*, 19(3):506–532.

Sorrel, M. A., Yigit, H. D., and Kaplan, K. (2018c). CD-CAT implementation of the proportional reasoning assessment. *Manuscript submitted for publication*.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement*, 37(1):58–75.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(3):337–350.

Tatsuoka, C. (2005). Corrigendum: Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(2):465–467.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of educational measurement*, 20(4):345–354.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. *Diagnostic monitoring of skill and knowledge acquisition*, pages 453–488.

Team, R. C. (2016). R: A language and environment for statistical computing. r foundation for statistical computing, Vienna, Austria. 2014.

Templin, J. (2006). CDM user's guide. *Unpublished manuscript*.

Templin, J. L. and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods*, 11(3):287.

Tjoe, H. and de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: an application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal*, 26(2):237–255.

Tu, D., Gao, X., Wang, D., and Cai, Y. (2017). A new measurement of internet addiction using diagnostic classification models. *Frontiers in psychology*, 8:1768.

von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series*, 2005(2).

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement*, 73(6):1017–1035.

Wang, C., Shu, Z., Shang, Z., and Xu, G. (2015). Assessing item-level fit for the DINA model. *Applied Psychological Measurement*, 39(7):525–538.

Wang, S., Yang, Y., Culpepper, S. A., and Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: A higher-order, hidden Markov model with covariates. *Journal of Educational and Behavioral Statistics*, page 1076998617719727.

Weekley, J. A., Hawkes, B., Guenole, N., and Ployhart, R. E. (2015). Low-fidelity simulations. *Annu. Rev. Organ. Psychol. Organ. Behav.*, 2(1):295–322.

Weiss, D. J. and Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21(4):361–375.

Whetzel, D. L. and McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review*, 19(3):188–202.

Willse, J. T. (2014). *CTT: Classical Test Theory Functions*. R package version 2.1.

Xu, G. and Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika*, 81(3):625–649.

Xu, X., Chang, H., and Douglas, J. (2003). A simulation study to compare CAT strategies for cognitive diagnosis. In *Annual meeting of the American Educational Research Association, Chicago*.

Xu, X. and von Davier, M. (2006). General diagnosis for NAEP proficiency data (ETS Research Rep. No. RR-06-08). *Princeton, NJ: Educational Testing Service*.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5(2):245–262.

Yi, Y.-S. (2017). Probing the relative importance of different attributes in l2 reading and listening comprehension items: An application of cognitive diagnostic models. *Language Testing*, 34(3):337–355.

Yigit, H. D., Sorrel, M. A., and de la Torre, J. (2018). Computerized adaptive testing for multiple-choice cognitive diagnosis models. *Manuscript submitted for publication*.